Confidence Sets for Statistical Classification (II): Exact Confidence Sets

W. Liu[1], F. Bretz[2], A.J. Hayter[3]

[1]S3RI and School of Mathematics

University of Southampton, SO17 1BJ, UK

[2]Novartis Pharma AG

Basel, 4002, Switzerland

[3]Dept of Statistics and Operations Technology

University of Denver, USA

Nov 2019

Abstract

Classification has applications in a wide range of fields including medicine, engineering, computer science and social sciences among others. Liu *et al.* (2019) proposed a confidence-set-based classifier that classifies a future object into a single class only when there is enough evidence to warrant this, and into several classes otherwise. By allowing classification of an object into possibly more than one class, this classifier guarantees a pre-specified proportion of correct classification among all future objects. However, the classifier uses a conservative critical constant. In this paper we show how to determine the exact critical constant in applications where prior knowledge about the proportions of the future objects from each class is available. As the exact critical constant is smaller than the conservative critical constant given in Liu *et al.* (2019), the classifier using the exact critical constant is better than the classifier in Liu *et al.* (2019) as expected. An example is provided to illustrate the method.

Keywords: Classification; Confidence level; Confidence set; Coverage frequency; Statistical inference.

1 Introduction

Classification has applications in a wide range of fields including medicine, engineering, computer science and social sciences among others. For overviews, the reader is referred to the books Webb and Copsey (2011), Flach (2012), Theodoridis and Koutroumbas (2009), Piegorsch (2015), and Hastie *et al.* (2017). In the recent paper, Liu *et al.* (2019) proposed a new classifier based on confidence sets. It constructs a confidence set for the the unknown parameter c, the true class of each future object, and classifies the object as belonging to the set of classes given by the confidence set. Hence this approach classifies a future object into a single class only when there is enough evidence to warrant this, and into several classes otherwise. By allowing classification of an object into potentially more than one class, this classifier guarantees a pre-specified proportion of correct classification among all future objects with a pre-specified confidence γ about the randomness in the training data based on which the classifier is constructed.

However, the classifier of Liu *et al.* (2019) uses a conservative critical constant λ and so the resultant confidence sets may be larger than necessary. The purpose of this paper is to determine the exact critical constant λ and therefore to improve the classifier of Liu *et al.* (2019) in situations where one has prior knowledge about the proportions of the (infinite) future objects belonging to the k possible classes.

The layout of the paper is as follows. Section 2 gives a very brief review of the classifier of Liu *et al.* (2019), and then considers the determination of the exact critical constant λ under the additional knowledge/assumption given above. An illustrative example is given in Section

3 to demonstrate the advantage of the improved classifier proposed in this paper when the additional assumption holds. Section 4 contains conclusions and discussions. Finally some mathematical details are provided in the appendix. As the same setting and notation as in Liu *et al.* (2019) are used, it is recommended to read this paper in conjunction with Liu *et al.* (2019).

2 Methodology

2.1 Methodology

Let the p-dimensional data vector $\mathbf{x}_l = (x_{l1}, \ldots, x_{lp})^T$ denote the feature measurement on an object from the lth class, which has multivariate normal distribution $N(\boldsymbol{\mu}_l, \Sigma_l)$, $l = 1, \ldots, k$; here k denotes the total number of classes which is a known number. The available training data set is given by $\mathcal{T} = \{\mathbf{x}_{l1}, \ldots, \mathbf{x}_{ln_l}; l = 1, \ldots, k\}$, where $\mathbf{x}_{l1}, \ldots, \mathbf{x}_{ln_l}$ are i.i.d. observations from the lth class with distribution $N(\boldsymbol{\mu}_l, \Sigma_l)$, $l = 1, \ldots, k$. The classification problem is to make inference about c, the true class of a future object, based on the feature measurement $\mathbf{y} = (y_1, \ldots, y_p)^T$ observed on the object, which is only known to belong to one of the k classes and so follows one of the k multivariate normal distributions. In statistical terminology, c is the unknown parameter of interest that takes a possible value in the simple parameter space $C = \{1, \ldots, k\}$. We emphasize that c is treated as non-random in both Liu *et al.* (2019) and here.

A classifier that classifies an object with measurement $\mathbf{y}$ into one single class in $C = \{1, \ldots, k\}$

can be regarded as a point estimator of c. The classifier of Liu *et al.* (2019) provides a set $\mathcal{C}_{\mathcal{T}}(\mathbf{y}) \subseteq C$ as plausible values of c. Depending on $\mathbf{y}$ and the training data set $\mathcal{T}$, $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ may contain only a single value, in which case $\mathbf{y}$ is classified into one single class given by $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$. When $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ contains more than one value in C, $\mathbf{y}$ is classified as possibly belonging to the several classes given by $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$. Hence, in statistical terms, the classifier uses the confidence set approach. The inherent advantage of the confidence set approach over the point estimation approach is the guaranteed $1-\alpha$ proportion of confidence sets that contain the true classes.

Specifically, the set $\mathcal{C}_{\mathcal{T}}(\mathbf{y}) \subseteq C$ is constructed in Liu *et al.* (2019) as

$$\mathcal{C}_{\mathcal{T}}(\mathbf{y}) = \left\{ l \in C : \ (\mathbf{y} - \hat{\boldsymbol{\mu}}_l)^T \hat{\Sigma}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_l) \leq \lambda \right\}, \tag{1}$$

where $\hat{\boldsymbol{\mu}}_l = \frac{1}{n_l} \sum_{m=1}^{n_l} \mathbf{x}_{lm}$ and $\hat{\Sigma}_l = \frac{1}{n_l - 1} \sum_{m=1}^{n_l} (\mathbf{x}_{lm} - \hat{\boldsymbol{\mu}}_l)(\mathbf{x}_{lm} - \hat{\boldsymbol{\mu}}_l)^T$, $l = 1, \ldots, k$, are respectively the usual estimators of the unknown $\boldsymbol{\mu}_l$ and Σ_l based on the training data set $\mathcal{T} = \{\mathbf{x}_{l1}, \ldots, \mathbf{x}_{ln_l}; l = 1, \ldots, k\}$, and λ is a suitably chosen critical constant whose determination is considered next. The intuition behind the definition of $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ in (1) is that a future object $\mathbf{y}$ is likely to be from class l if and only if $(\mathbf{y} - \hat{\boldsymbol{\mu}}_l)^T \hat{\Sigma}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_l) \leq \lambda$.

Note that the proportion of the future confidence sets $\mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)$ $(j = 1, 2, \ldots)$ that include the true classes c_j of $\mathbf{y}_j$ $(j = 1, 2, \ldots)$ is given by $\liminf_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} I_{\{c_j \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}}$. So it is desirable that

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} I_{\{c_j \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}} \geq 1 - \alpha \tag{2}$$

where $1 - \alpha$ is a pre-specified large (close to 1) proportion, 0.95 say. While the constraint in (2) is difficult to deal with, it is shown in Liu *et al.* (2019) that a sufficient condition for guaranteeing (2) is

$$\inf_{c_j \in C} E_{\mathbf{y}_j|\mathcal{T}} I_{\{c_j \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}} \geq 1 - \alpha \tag{3}$$

where $E_{\mathbf{y}_j|\mathcal{T}}$ denotes the conditional expectation with respect to the random variable $\mathbf{y}_j$ conditioning on the training data set $\mathcal{T}$ (or, equivalently, $\{(\hat{\boldsymbol{\mu}}_1, \hat{\Sigma}_1), \ldots, (\hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)\}$).

Since the value of the expression on the left hand side of the inequality in (3) (and in (2) as well) depends on $\mathcal{T}$ and $\mathcal{T}$ is random, the inequality in (3) cannot be guaranteed for each observed $\mathcal{T}$. We therefore guarantee (3) with a large (close to 1) probability γ with respect to the randomness in $\mathcal{T}$:

$$P_{\mathcal{T}} \left\{ \inf_{c_j \in C} E_{\mathbf{y}_j|\mathcal{T}} I_{\{c_j \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}} \geq 1 - \alpha \right\} = \gamma, \tag{4}$$

which in turn guarantees that

$$P_{\mathcal{T}} \left\{ \liminf_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} I_{\{c_j \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}} \geq 1 - \alpha \right\} \geq \gamma. \tag{5}$$

Computer code in R is provided in Liu *et al.* (2019) to compute the λ that solves the equation in (4), which allows the confidence sets $\mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)$ in (1) to be constructed for each future object.

The interpretation of the expressions in (5) and (6) below is that, based on one observed training data set $\mathcal{T}$, one constructs confidence sets $\mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)$ for the c_j's of all future $\mathbf{y}_j$ ($j = 1, 2, \cdots$) and claims that at least $1 - \alpha$ proportion of these confidence sets do contain the true c_j's. Then we are γ confident with respect to the randomness in the training data set $\mathcal{T}$ that the claim is correct.

A natural question is how to find the exact critical constant λ that solves the equation

$$P_{\mathcal{T}} \left\{ \liminf_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} I_{\{c_j \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}} \geq 1 - \alpha \right\} = \gamma \tag{6}$$

which is an improvement to the conservative λ that solves the equation in (4) as given in Liu *et al.* (2019). Next we show how to find the exact critical constant λ under an additional assumption which is satisfied in some applications.

Assume that, among the N future objects that need to be classified, N_l objects are actually from the lth class with the distribution $N(\boldsymbol{\mu}_l, \Sigma_l)$, $l = 1, \ldots, k$. The additional assumption we make is that

$$\lim_{N \to \infty} \frac{N_l}{N} = r_l, \quad l = 1, \ldots, k \tag{7}$$

where the r_l's are assumed to be known constants in the interval $[0, 1]$. Intuitively this assumption means that we know the proportions of the future objects that belong to each of the k classes, even though we do not know the true class of each individual future object.

The assumption in (7) is reasonable in some applications. For example, when screening for a particular disease among a specific population for preventive purpose, there are $k = 2$ classes: having the disease ($l = 1$) or not having the disease ($l = 2$). If we know the prevalence of the disease, d, in the overall population then $r_1 = d$ and $r_2 = 1 - d$, even though we do not know whether an individual subject has the disease or not.

It is shown in the Appendix that, under the assumption in (7), the equation in (6) is equivalent to

$$P_{\mathbf{u}_l, \{\mathbf{v}_{lm}\}} \left\{ \sum_{l=1}^{k} r_l P_{\mathbf{w}_l \,|\, \mathbf{u}_l, \{\mathbf{v}_{lm}\}} \left\{ (\mathbf{w}_l - \mathbf{u}_l)^T \left(\frac{1}{n_l - 1} \sum_{m=1}^{n_l - 1} \mathbf{v}_{lm} \mathbf{v}_{lm}^T \right)^{-1} (\mathbf{w}_l - \mathbf{u}_l) \leq \lambda \right\} \geq 1 - \alpha \right\} = \gamma \tag{8}$$

where

$$\mathbf{w}_l \sim N(\mathbf{0}, I_p), \quad \mathbf{u}_l \sim N(\mathbf{0}, I_p/n_l), \quad \mathbf{v}_{lm} \sim N(\mathbf{0}, I_p), \, m = 1, \cdots, n_l - 1 \tag{9}$$

and all the $\mathbf{w}_l$'s, $\mathbf{u}_l$'s and $\mathbf{v}_{lm}$'s are independent, $P_{\mathbf{w}_l \,|\, \mathbf{u}_l, \{\mathbf{v}_{lm}\}}\{\cdot\}$ denotes the conditional probability about $\mathbf{w}_l$ conditioning on $(\mathbf{u}_l, \{\mathbf{v}_{lm}\})$, and $P_{\mathbf{u}_l, \{\mathbf{v}_{lm}\}}\{\cdot\}$ denotes the probability about $(\mathbf{u}_l, \{\mathbf{v}_{lm}\})$.

2.2 Algorithm for computing the exact λ

We now consider how to compute the critical constant λ that solves the equation in (8). Similar to Liu *et al.* (2019), this is accomplished by simulation in the following way. From the distributions given in (9), in the sth repeat of simulation, $s = 1, \ldots, S$, generate independent

$$\mathbf{u}_l^s \sim N(\mathbf{0}, I_p/n_l)\,, \quad \mathbf{v}_{l1}^s, \ldots, \mathbf{v}_{l(n_l-1)}^s \sim N(\mathbf{0}, I_p)\,; \quad l = 1, \ldots, k.$$

and find the $\lambda = \lambda_s$ so that

$$\sum_{l=1}^{k} r_l P_{\mathbf{w}_l \,|\, \mathbf{u}_l^s, \{\mathbf{v}_{lm}^s\}} \left\{ (\mathbf{w}_l - \mathbf{u}_l^s)^T \left(\frac{1}{n_l - 1} \sum_{m=1}^{n_l-1} \mathbf{v}_{lm}^s \mathbf{v}_{lm}^{s\ T} \right)^{-1} (\mathbf{w}_l - \mathbf{u}_l^s) \le \lambda_s \right\} = 1 - \alpha. \qquad (10)$$

Repeat this S times to get $\lambda_1, \ldots, \lambda_S$ and order these as $\lambda_{[1]} \le \ldots \le \lambda_{[S]}$. It is well known (cf. Serfling, 1980) that $\lambda_{[\gamma S]}$ converges to the required critical constant λ with probability one as $S \to \infty$. Hence $\lambda_{[\gamma S]}$ is used as the required critical constant λ for a large S value, 10,000 say.

To find the λ_s in (10) for each s, we use simulation in the following way. Generate independent random vectors $\{\mathbf{w}_{lq} : q = 1, \ldots, Q; l = 1, \ldots, k\}$ from $N(\mathbf{0}, I_p)$, where Q is the number of simulations for finding λ_s. For each given value of $\lambda_s > 0$, the expression on the left-side of the equation in (10) can be computed by approximating each of the k probabilities involved using the corresponding proportions out of the Q simulations. It is also clear that this expression is monotone increasing in λ_s. Hence the λ_s that solves the equation in (10) can be found by using a searching algorithm; for example, the bi-section method is used in our R code. To approximate reasonably accurately the probabilities with the proportions, a large Q value, 10,000 say, should be used.

It is noteworthy from (8) and (9) that λ depends only on $\gamma, \alpha, p, k, n_1, \ldots, n_k, r_1, \ldots, r_k$ (and the numbers of simulations S and Q which determine the numerical accuracy of λ due to simulation randomness). It is also worth emphasizing that only one λ needs to be computed based on the observed training dataset $\mathcal{T}$ which is then used for constructing the confidence sets $\mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)$ and classifying accordingly all future objects.

It is expected that larger values of S and Q will produce more accurate λ value, one can use the method discussed in Liu *et al.* (2019) to assess how the accuracy of λ depends on the values of S and Q. As in Liu *et al.* (2019), it is recommended to set $S = 10,000$ and $Q = 10,000$ for reasonable computation time and accuracy of λ due to simulation randomness.

3 An illustrative example

As in Liu *et al.* (2019), the famous `iris` data set introduced by Fisher (1936) is used in this section to illustrate the method proposed in this paper. The data set contains $k = 3$ classes representing the three species/classes of Iris flowers (1=setosa, 2=versicolor, 3=virginica), and has $n_i = 50$ observations from each class in $\mathcal{T}$. Each observation gives the measurements (in centimetres) of the four variables: sepal length and width, and petal length and width.

We focus on the case that only the first two measurements, sepal length and width, are used for classification in order to easily illustrate the method since the acceptance sets $\mathcal{A}_l = \left\{ \mathbf{y} \in R^p : \ (\mathbf{y} - \hat{\boldsymbol{\mu}}_l)^T \hat{\Sigma}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_l) \le \lambda \right\}$, $l = 1, 2, 3$ are two-dimensional and so can be easily plotted in this case. Based on the fifty observations on $p = 2$ measurements from each of the three classes, the $\hat{\boldsymbol{\mu}}_l$ and $\hat{\Sigma}_l$ are given in Liu *et al.* (2019).

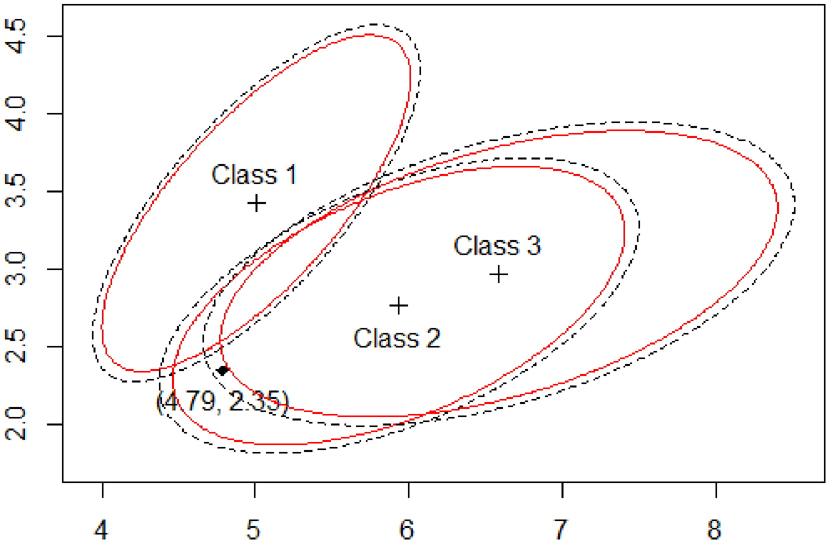

Figure 1: The exact (solid) and conservative (dotted) acceptance sets for the three classes

For $\alpha = 5\%$ and $\gamma = 95\%$, the critical constant λ that solves the equation in (4) is computed in Liu *et al.* (2019) to be $\lambda_{con} = 9.175$ using $S = 10,000$ and $Q = 10,000$. The corresponding acceptance sets, based on which the confidence set $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ in (1) can be constructed directly (cf. Liu *et al.*, 2019), are given by

$$\mathcal{A}_l^{con} = \left\{ \mathbf{y} \in R^p : \ (\mathbf{y} - \hat{\boldsymbol{\mu}}_l)^T \hat{\Sigma}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_l) \le \lambda_{con} \right\}, \ l = 1, 2, 3$$

and plotted in Figure 1 by the dotted ellipsoidal region centred at $\hat{\boldsymbol{\mu}}_l$, marked by '+'.

Now assume that we have the knowledge about the proportions of the three species among all the Iris flowers (r_1, r_2, r_3) and the Iris flowers that need to be classified reflect this composition. For the same $\alpha = 5\%$, $\gamma = 95\%$, $S = 10,000$ and $Q = 10,000$, and with, for example, $(r_1, r_2, r_3) = (0.3, 0.4, 0.3)$, the exact critical constant λ that solves the equation in (6) is computed by our R program to be $\lambda_{exa} = 7.737$. As expected, λ_{exa} is smaller than λ_{con} and, as a result, the corresponding confidence set $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ in (1) with $\lambda = \lambda_{exa}$ and acceptance sets $\mathcal{A}_l^{exa} = \left\{ \mathbf{y} \in R^p : \ (\mathbf{y} - \hat{\boldsymbol{\mu}}_l)^T \hat{\Sigma}_l^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}_l) \leq \lambda_{exa} \right\}$, $l = 1, 2, 3$, are also smaller than the $\mathcal{A}_l^{con}$ given in Liu *et al.* (2019).

The acceptance sets $\mathcal{A}_l^{exa}$, $l = 1, 2, 3$ are plotted in Figure 1 by the solid ellipsoidal regions. For example, if a future object has $\mathbf{y} = (4.79, 2.35)$, marked by a solid dot in Figure 1, then the conservative confidence set of Liu *et al.* (2019) classifies the object as from classes 2 and 3 since this $\mathbf{y}$ belongs to both $\mathcal{A}_2^{con}$ and $\mathcal{A}_3^{con}$. But the new exact confidence set of this paper classifies the object as from class 2 only since this $\mathbf{y}$ belongs to $\mathcal{A}_2^{exa}$ but not $\mathcal{A}_1^{exa}$ or $\mathcal{A}_3^{exa}$. This demonstrates the advantage of the new confidence set using λ_{exa} in this paper over the conservative confidence set using λ_{con} in Liu *et al.* (2019). We have also computed the value of λ_{exa} for several other given (r_1, r_2, r_3). For example, $\lambda_{exa} = 7.706$ for $(r_1, r_2, r_3) = (1/3, 1/3, 1/3)$, $\lambda_{exa} = 7.865$ for $(r_1, r_2, r_3) = (0.1, 0.45, 0.45)$, and $\lambda_{exa} = 8.019$ for $(r_1, r_2, r_3) = (0.1, 0.7, 0.2)$. The conservative $\lambda_{con} = 9.175$ is considerably, ranging from 14% to 19%, larger than these λ_{exa} values.

One can download from `http://www.personal.soton.ac.uk/wl/Classification/` the R computer program `ExactConfidenceSetClassifier.R` that implements this simulation method of computing the critical constant λ_{exa}. The computation of one λ_{exa} using $(S, Q) =$

$(10,000,\ 10,000)$ takes about 13 hours on an ordinary Window's PC (Core(TM2) Duo CPU P8400@2.26GHz).

It must be emphasized though the new confidence set is valid only if the assumption in (7) is true. If the assumption does not holds, then the conservative confidence set of Liu *et al.* (2019) should be used in order for the statement in (5) to hold.

4 Conclusions

The probability statement in (5) allows that the confidence sets in Liu *et al.* (2019) have the nice interpretation that, with confidence level γ about the randomness in the training data set $\mathcal{T}$, at least $1-\alpha$ proportion of the confidence sets $\mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)$, $j = 1, 2, \ldots$ contain the true classes c_j, $j = 1, 2, \ldots$ of the future objects $\mathbf{y}_j$, $j = 1, 2, \ldots$. However, the confidence set given in Liu *et al.* (2019) is conservative in that the λ in the confidence set in (1) is computed to solve the equation in (4), which implies the constraint in (5). This paper considers how to compute the λ in the confidence set in (1) so that the probability in (5) is equal to γ, i.e. from the equation in (6). The confidence sets using the λ that solves the equation in (6) have the confidence level equal to γ and so are exact. We show that this can be accomplished under the extra assumption given in (7), which may be sensible in some applications.

As the λ_{exa} that solves the equation in (6) is smaller than the λ_{con} that solves the equation in (4) used in Liu *et al.* (2019), the new confidence sets are smaller and so better than the confidence sets given in Liu *et al.* (2019).

One wonders whether there are other sensible assumptions that allow the λ to be solved from the equation (6). This warrants further research.

If $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ for a future object $\mathbf{y}$ is empty then, since $\mathbf{y}$ must be from one of the k classes, $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ can be augmented to include the class that has the largest posterior probability using the naive Bayesian classifier as in Liu *et al.* (2019). The probability statement in (5) clearly holds under this augmentation to $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ only when $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ is empty.

There are applications in which information about the proportions r_l would be known with uncertainty. For example, the training set may be a representative sample from the population and as such the proportion of each class can be estimated, or the proportions might have been estimated by a previous independent auxiliary dataset. If one replaces the r_l's in the expression in (8) by these estimates then the λ solved from the equation in (8) will depend on these estimates and so be random. As a result, the probability statement in (5) is no longer valid. How to deal with these applications warrants further research.

Finally, the classifier of Liu *et al.* (2019) is developed from the idea of Lieberman *et al.* (1963, 1967). The same idea has also been used in, for example, Mee *et al.* (1991), Han *et al.* (2016), Liu *et al.* (2016) and Peng *et al.* (2019) which all use conservative critical constants as in Liu *et al.* (2019). The idea of this paper can be applied to all these works to compute exact critical constants under suitable extra assumptions.

5 Appendix: Mathematical details

In this appendix we show the equivalence of the equations in (6) and (8) under the assumption in (7). Note first the well known fact (cf. Anderson, 2003) that $\hat{\boldsymbol{\mu}}_l \sim N(\boldsymbol{\mu}_l, \Sigma_l/n_l)$, $(n_l - 1)\hat{\Sigma}_l = \sum_{m=1}^{n_l-1} \mathbf{z}_{lm}\mathbf{z}_{lm}^T$ with $\mathbf{z}_{l1}, \ldots, \mathbf{z}_{l(n_l-1)}$ being i.i.d. $N(\mathbf{0}, \Sigma_l)$ random vectors independent of $\hat{\boldsymbol{\mu}}_l$.

Among the N future objects that need to be classified, let N_l be the number of objects actually from the lth class with the feature measurements denoted as $\mathbf{y}_{l1}, \ldots, \mathbf{y}_{lN_l}$, $l = 1, \ldots, k$. Clearly we have $N_1 + \cdots + N_k = N$ and

$$\begin{aligned} & \liminf_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} I_{\{c_j \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}} \\ = & \liminf_{N\to\infty} \frac{1}{N} \sum_{l=1}^{k} \sum_{i=1}^{N_l} I_{\{c_l \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_{li})\}} \\ = & \liminf_{N\to\infty} \sum_{l=1}^{k} \frac{N_l}{N} \left(\frac{1}{N_l} \sum_{i=1}^{N_l} I_{\{c_l \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_{li})\}} \right). \end{aligned} \tag{11}$$

We have from the classical strong law of large numbers (cf. Chow and Teicher, 1978) that

$$\lim_{N_l\to\infty} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[I_{\{c_l \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_{li})\}} - E_{\mathbf{y}_{li}|\mathcal{T}} I_{\{c_l \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_{li})\}} \right] = 0, \tag{12}$$

in which the conditional expectation $E_{\mathbf{y}_{li}|\mathcal{T}}$ is used since all the confidence sets $\mathcal{C}_{\mathcal{T}}(\mathbf{y}_{li})$ $(i = 1, \ldots, N_l)$ use the same training data set $\mathcal{T}$. By noting that $\mathbf{y}_{li}, i = 1, \ldots, N_l$ are from the lth class and so have the same distribution $N(\boldsymbol{\mu}_l, \Sigma_l)$, we have from the definition of $\mathcal{C}_{\mathcal{T}}(\mathbf{y})$ in (1) that

$$\begin{aligned} & E_{\mathbf{y}_{li}|\mathcal{T}} I_{\{c_l \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_{li})\}} \\ = & P_{\mathbf{y}_{l1}|\mathcal{T}} \{c_l \in \mathcal{C}_{\mathcal{T}}(\mathbf{y}_{l1})\} \end{aligned}$$

$$
\begin{aligned}
&= P_{\mathbf{y}_{l1}|\mathcal{T}}\left\{(\mathbf{y}_{l1}-\hat{\boldsymbol{\mu}}_l)^T\hat{\Sigma}_l^{-1}(\mathbf{y}_{l1}-\hat{\boldsymbol{\mu}}_l)\le\lambda\right\} \\
&= P_{\mathbf{w}_l\,|\,\mathbf{u}_l,\{\mathbf{v}_{lm}\}}\left\{(\mathbf{w}_l-\mathbf{u}_l)^T\left(\frac{1}{n_l-1}\sum_{m=1}^{n_l-1}\mathbf{v}_{lm}\mathbf{v}_{lm}^T\right)^{-1}(\mathbf{w}_l-\mathbf{u}_l)\le\lambda\right\} \qquad (13)
\end{aligned}
$$

where

$$
\begin{aligned}
&\mathbf{w}_l = \Sigma_l^{-1/2}(\mathbf{y}_{l1}-\boldsymbol{\mu}_l)\sim N(\mathbf{0},I_p) \\
&\mathbf{u}_l = \Sigma_l^{-1/2}(\hat{\boldsymbol{\mu}}_l-\boldsymbol{\mu}_l)\sim N(\mathbf{0},I_p/n_l) \\
&\mathbf{v}_{lm} = \Sigma_l^{-1/2}\mathbf{z}_{lm}\sim N(\mathbf{0},I_p),\ m=1,\cdots,n_l-1
\end{aligned}
$$

with all the $\mathbf{w}_l$'s, $\mathbf{u}_l$'s and $\mathbf{v}_{lm}$'s being independent. Note that $\mathbf{w}_l$ depends on the future observation $\mathbf{y}_{l1}$ but not the training data set $\mathcal{T}$, while $\mathbf{u}_l$ and $\{\mathbf{v}_{lm}\}$ depend on the training data set $\mathcal{T}$ but not the future observations.

Combining the assumption in (7) and the expressions in (11), (12) and (13) gives

$$
\liminf_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}I_{\{c_j\in\mathcal{C}_{\mathcal{T}}(\mathbf{y}_j)\}} = \sum_{l=1}^{k}r_lP_{\mathbf{w}_l\,|\,\mathbf{u}_l,\{\mathbf{v}_{lm}\}}\left\{(\mathbf{w}_l-\mathbf{u}_l)^T\left(\frac{1}{n_l-1}\sum_{m=1}^{n_l-1}\mathbf{v}_{lm}\mathbf{v}_{lm}^T\right)^{-1}(\mathbf{w}_l-\mathbf{u}_l)\le\lambda\right\},
$$

from which the equivalence of the equations in (6) and (8) follows immediately.

Acknowledgements: We would like to thank the referees for critical and constructive comments on the earlier version of the paper.

References

[1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis, 3rd edition.* Wiley.

[2] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179-188.

[3] Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press

[4] Han, Y., Liu, W., Bretz, F., Wan, F., Yang, P. (2016). Statistical calibration and exact one-sided simultaneous tolerance intervals for polynomial regression. *Journal of Statistical Planning and Inference*, 168, 90–96.

[5] Hastie, T., Tibshirani, R. and Friedman, J. (2017) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. New York: Springer.

[6] Lieberman, G.J. and Miller, R.G., Jr. (1963). Simultaneous tolerance intervals in regression. *Biometrika*, 50, 155–168.

[7] Lieberman, G.J., Miller, R.G., Jr. and Hamilton, M.A. (1967). Simultaneous discrimination intervals in regression. *Biometrika*, 54, 133–145 (Correction: 58, 687).

[8] Liu, W., Bretz, F., Srimaneekarn, N., Peng, J. and Hayter, A.J. (2019). Confidence sets for statistical classification. *Stats*, 2(3), 332-346. https://doi.org/10.3390/stats2030024

[9] Liu, W., Han, Y., Bretz, F., Wan, F., Yang, P. (2016). Counting by weighing: know your numbers with confidence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65, 641–648.

[10] Mee, R.W., Eberhardt, K.R. and Reeve, C.P. (1991). Calibration and simultaneous tolerance intervals for regression. *Technometrics*, 33, 211–219.

[11] Peng, J., Liu, W., Bretz, F. and Hayter, A.J. (2019). Counting by weighing: two-sided confidence intervals. *Journal of Applied Statistics*, 46(2), 262-271.

[12] Piegorsch, W.W. (2015). *Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery*. Wiley.

[13] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

[14] Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition, 4th edition*. Academic Press.

[15] Webb, A.R. and Copsey, K.D. (2011). *Statistical Pattern Recognition, 3rd edition*. Wiley