Journal of Data Science 0 (0), 1–5
July 2022
DOI: 0000
Research Article

Supplementary document for "A two-stage classification for dealing with unseen clusters in the testing data"

Jung Wun Lee[1] and Ofer Harel[2,*]

[1]Department of Biostatistics, Harvard University, Boston, MA, 02115, USA
[2]Department of Statistics, University of Connecticut, Storrs, CT, 06269, USA

Proof of Theorem 2.1

Proof. The conditional expectation of $\psi_\alpha(\mathbf{x})$ given $\omega_\mathbf{x} \in \Omega_K$ can be obtained as follows.

$$\begin{aligned} E(\psi_\alpha(x) \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) &= P(\psi_\alpha^{(1)}(\mathbf{x}) = 1, \psi_\alpha^{(2)}(\mathbf{x}) = 1, \ldots, \psi_\alpha^{(K)}(\mathbf{x}) = 1 \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \\ &= P(\bigcap_{k=1}^{K} \{\psi_\alpha^{(k)}(\mathbf{x}) = 1\} \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \\ &\leqslant \min_k \left\{ P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \right\} \qquad (1) \\ &= \min_k \left\{ \sum_{m=1}^{K} P(\psi_\alpha^{(k)}(\mathbf{x}) = 1, \mathbf{x} \in A_m \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \right\}. \end{aligned}$$

Here, $P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) = \sum_{m=1}^{K} P(\psi_\alpha^{(k)}(\mathbf{x}) = 1, \mathbf{x} \in A_m \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j)$ because A_k are disjoint sets. Next, let $p_m = P(\mathbf{x} \in A_m \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j)$ be the proportion of A_m among $[A_1, \ldots, A_K]$ thus $\sum_{m=1}^{K} p_m = 1$. Then, we have the following decomposition.

$$\begin{aligned} P(\psi_\alpha^{(k)}(\mathbf{x}) = 1, \mathbf{x} \in A_m \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) &= P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in A_m, \mathbf{x} \in \bigcup_{j=1}^{K} A_j) P(\mathbf{x} \in A_m \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \\ &= P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in A_m) P(\mathbf{x} \in A_m \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \\ &= P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in A_m) p_m. \qquad (2) \end{aligned}$$

Finally, the type 1 error $E(\psi_\alpha(\mathbf{x}) \mid \omega_\mathbf{x} \in \Omega_K)$ has upper bound as follows.

*Corresponding author. Email: jwlee@hsph.harvard.edu or ofer.harel@uconn.edu.

$$E(\psi_\alpha(\mathbf{x}) \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \leqslant \min_k \left\{ \sum_{m=1}^{K} P(\psi_\alpha^{(k)}(\mathbf{x}) = 1, \mathbf{x} \in A_m \mid \mathbf{x} \in \bigcup_{j=1}^{K} A_j) \right\}$$
$$\leqslant \min_k \left\{ \sum_{m=1}^{K} P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in A_m) p_m \right\}. \quad (3)$$

Here, $P(\psi_\alpha^{(m)}(\mathbf{x}) = 1 \mid \mathbf{x} \in A_m) = \alpha$ by definition, and $P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in A_m, k \neq m)$ is a probability of correct decision and thus is clearly bounded above by 1. Now, Eq. (3) can be simplified as follows.

$$E(\psi_\alpha(\mathbf{x}) \mid \omega_{\mathbf{x}} \in \Omega_K) \leqslant \min_k \left\{ \sum_{m=1}^{K} P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \in A_m) p_m \right\}$$
$$\leqslant \min_k \left\{ \sum_{m=1}^{K} \alpha p_m, \sum_{m=1}^{K} p_m, \ldots, \sum_{m=1}^{K} p_m \right\} = \alpha. \quad (4)$$

As shown in Eq. (4), a size of the test $\psi_\alpha(\mathbf{x})$ is less than α. □

Proof of Theorem 2.2

Proof. The power of the test $\psi_\alpha(\mathbf{x})$ is the probability of rejecting the null hypothesis $H_0 : \omega_{\mathbf{x}} \in \Omega_K$ under $H_1 : \omega_{\mathbf{x}} \notin \Omega_K$. To achieve $\psi_\alpha(\mathbf{x}) = 1$, all K sub-tests $\psi_\alpha^{(k)}(\mathbf{x}), k = 1, \ldots, K$ should have value 1. Consequently, the power of the test $\psi_\alpha(\mathbf{x})$ can be written as follows.

$$P(\psi_\alpha(\mathbf{x}) = 1 \mid \mathbf{x} \notin \bigcup_{j=1}^{K} A_j) = P(\psi_\alpha^{(1)}(\mathbf{x}) = 1, \ldots, \psi_\alpha^{(K)}(\mathbf{x}) = 1 \mid \mathbf{x} \notin \bigcup_{j=1}^{K} A_j)$$
$$\leqslant \min_k \left\{ P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \mathbf{x} \notin \bigcup_{j=1}^{K} A_j) \right\}. \quad (5)$$

As shown in Eq. (5), the power of the test is bounded above the smallest conditional probability of rejecting $H_0 : \omega_{\mathbf{x}} = k$ given that $\omega_{\mathbf{x}} \notin \Omega_K$. □

Proof of Theorem 2.2

Proof. Let A be the event of correct classification of a testing observation, and A^c denotes an event of misclassification. Also, and D be the event of the unseen-cluster problem. Also, let $P_0(A)$ and $P_1(A)$ be the probability of correct classification by M_0 and M_1, respectively. Then, we can write $\zeta_0^{(\delta)}$ as a function of $[\delta, \beta_0]$ as follows.

$$\begin{aligned} \zeta_0^{(\delta)} &= P_0(A) = P_0(A \cap D) + P_0(A \cap D^c) \\ &= P_0(A \mid D) P_0(D) + P_0(A \mid D^c) P_0(D^c) \\ &= 0 \times \delta + \beta_0(1 - \delta) = \beta_0(1 - \delta). \end{aligned} \quad (6)$$

Here, $P_0(A \mid D) = 0$ because a conventional method can never correctly classify a testing subject if it is sampled from unseen-cluster. Also, we have $P_0(D^c) = P_1(D^c) = 1 - \delta$. Now, let B be the event of diagnosis of the unseen-cluster. Similarly, $\zeta_1^{(\delta)}$ can be written as follows.

$$\begin{aligned}
\zeta_1^{(\delta)} &= P_1(A) = P_1(A \cap D) + P_1(A \cap D^c) = P_1(A \mid D)P(D) + P_1(A \mid D^c)P(D^c) \\
&= \{P_1(A \cap B \mid D) + P_1(A \cap B^c \mid D)\} P_1(D) \\
&+ \{P_1(A \cap B \mid D) + P_1(A \cap B^c \mid D^c)\} P_1(D^c) \\
&= P_1(A \cap B \mid D)P_1(D) + P_1(A \cap B^c \mid D^c)P_1(D^c) \\
&= P_1(A \mid B, D)P_1(B \mid D)P_1(D) + P_1(A \mid B^c, D^c)P_1(B^c \mid D^c)P_1(D^c) \\
&= \delta\beta_1 + (1 - \delta)(1 - \alpha)\beta_0. \qquad (7)
\end{aligned}$$

In Eq. (7), we have $P_1(A \cap B^c \mid D) = P_1(A \cap B \mid D^c) = 0$ and $P_1(A \mid B, D) = 1$. This is because a correct classification is impossible if a unseen-cluster is not detected (i.e., $P_1(A \mid D) = 0$). Similarly, an event $B \mid D^c$ denotes a false discovery event for a unseen-cluster where a correct classification is impossible, thus $P_1(B \mid D^c) = 0$. In addition, we have $P_1(A \mid B, D) = 1$ because two-stage classification successfully classifies a testing observation as long as the test correctly identifies it as *unclassified*. Similarly, we have $P_1(A \cap B^c \mid D) = P_1(A \cap B \mid D^c) = 0$ because the false discovery or failure of diagnosis of unseen-cluster problem makes the probability of correct classification to be 0. Subtracting Eq. (6) from Eq. (7), the theorem is proved. □

Numerical studies when K is unspecified

Numerical studies in Section 3.4 and the following results are based on simulation studies where the true number of clusters in training data is correctly specified, which can be misspecified in practice. In this section, we investigate the performances of the proposed methods without using the true value of K. Namely, when establishing the diagnostic test $\psi_\alpha(\mathbf{x})$, we do not employ the true training cluster number $K = 4$. Instead, we estimate the Gaussian mixture model with different numbers of clusters ranging from 2 to 9 and use the model with the lowest BIC for establishing $\psi_\alpha(\mathbf{x})$. We expect that such a process will show comparable performances with results using true K values when the cluster separation is strong, and thus, BIC correctly identifies the true value of K. On the other hand, when underlying training cluster separations are weak, BIC often underestimates K, and thus, the performances of the following diagnostic test and two-stage classification may decrease. For the rest of the subsection, we investigate type 1 error rates and power of the diagnostic tests, and prediction accuracies of two-stage classification without using K.

Table 1 illustrates the average type 1 error of the diagnostic test when the training cluster number is unspecified and estimated via BIC. Similar to the scenarios with true K, the proposed method controls the type 1 error rates under all scenarios, while other competing diagnostic tests overuse the type 1 error rates.

Table 2 illustrates the average power of the diagnostic test when the training cluster number is unspecified. As expected, all diagnostic tests illustrate similar performances when the training clusters are strongly separated (that is, when $\mu = 4$ and $[\mu, \sigma] = [3.0, 0.5]$. When the training cluster separations are weak, such as $\mu = 2.0$, the average power of all diagnostic tests decreases compared to the scenarios with true K used (See Table 2 in the main paper). If embedded in two-stage classification, this may yield a noticeable decrease in prediction accuracies.

Table 1: Average Type 1 errors of diagnostic tests when K is unspecified and estimated via BIC.

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$\sigma = 0.5$	0.053	0.064	0.053	0.053	0.062	0.060	0.063
	$\sigma = 1.0$	0.045	0.063	0.052	0.053	0.061	0.059	0.063
	$\sigma = 1.2$	0.043	0.061	0.053	0.053	0.059	0.059	0.060
$\mu = 3.0$	$\sigma = 0.5$	0.046	0.064	0.053	0.053	0.062	0.059	0.063
	$\sigma = 1.0$	0.042	0.058	0.052	0.053	0.057	0.059	0.058
	$\sigma = 1.2$	0.043	0.059	0.053	0.053	0.057	0.059	0.058
$\mu = 2.0$	$\sigma = 0.5$	0.047	0.058	0.052	0.053	0.057	0.058	0.058
	$\sigma = 1.0$	0.046	0.059	0.052	0.053	0.057	0.058	0.058
	$\sigma = 1.2$	0.047	0.058	0.052	0.053	0.057	0.058	0.058

Table 2: Average power of diagnostic tests when K is unspecified and estimated via BIC.

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$\sigma = 0.5$	0.962	0.967	0.869	0.940	0.966	0.955	0.966
	$\sigma = 1.0$	0.669	0.705	0.438	0.566	0.696	0.638	0.704
	$\sigma = 1.2$	0.566	0.607	0.321	0.416	0.548	0.542	0.605
$\mu = 3.0$	$\sigma = 0.5$	0.735	0.766	0.514	0.649	0.766	0.705	0.764
	$\sigma = 1.0$	0.379	0.414	0.182	0.219	0.278	0.371	0.412
	$\sigma = 1.2$	0.314	0.347	0.157	0.185	0.226	0.313	0.345
$\mu = 2.0$	$\sigma = 0.5$	0.336	0.369	0.165	0.194	0.240	0.332	0.367
	$\sigma = 1.0$	0.178	0.201	0.107	0.117	0.146	0.189	0.200
	$\sigma = 1.2$	0.156	0.176	0.098	0.106	0.136	0.167	0.175

Finally, Table 3 illustrates the average prediction accuracies of the two-stage classification when the proportion of unseen-cluster is 10% in the testing data, and the number of training clusters in the embedded diagnostic test is unspecified and estimated via BIC. Similar to the power in Table 2, prediction accuracies of two-stage classification methods are lower than in the scenario with true K when the training cluster separation is weak, as shown in Table 5 in the main paper. This is because the power of the diagnostic tests decreases when K is underestimated, triggering prediction accuracies to decrease accordingly.

We conclude that the loss of power and prediction accuracies due to underestimating K due to overlapping clusters is not an inherent disadvantage of our proposed method. Unfortunately, dealing with overlapping clusters is fundamentally challenging in the finite mixture model and, thus, a natural limitation of the multiclass-based novelty detection framework. In this sense, we emphasize the importance of correctly estimating the number of clusters in the training data. We refer to Xu et al. (2016) as a recent review on determining the number of clusters. The infinite-mixture model can be another interesting approach. Still, it is not a decision-free method because the hyperparameters of the Dirichlet-Process prior distribution on K should be carefully specified (Li et al., 2019).

Table 3: Average prediction accuracies of two-stage classifications at 10% unseen-clusters when K is unspecified and estimated via BIC.

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
	$\sigma = 0.5$	0.948	0.944	0.937	0.944	0.944	0.942	0.944
$\mu = 4.0$	$\sigma = 1.0$	0.850	0.848	0.822	0.835	0.848	0.840	0.848
	$\sigma = 1.2$	0.683	0.686	0.658	0.663	0.677	0.680	0.686
	$\sigma = 0.5$	0.877	0.874	0.850	0.864	0.874	0.866	0.874
$\mu = 3.0$	$\sigma = 1.0$	0.662	0.664	0.642	0.646	0.655	0.656	0.664
	$\sigma = 1.2$	0.633	0.635	0.616	0.620	0.627	0.628	0.635
	$\sigma = 0.5$	0.641	0.644	0.623	0.627	0.635	0.636	0.643
$\mu = 2.0$	$\sigma = 1.0$	0.552	0.555	0.543	0.545	0.552	0.550	0.555
	$\sigma = 1.2$	0.532	0.535	0.524	0.526	0.532	0.531	0.534

References

Li Y, Schofield E, Gönen M (2019). A tutorial on dirichlet process mixture modeling. *Journal of Mathematical Psychology*, 91: 128–144.

Xu S, Qiao X, Zhu L, Zhang Y, Xue C, Li L (2016). Reviews on determining the number of clusters. *Applied Mathematics & Information Sciences*, 10(4): 1493–1512.