arXiv:2304.03990v2 [cs.LG] 5 Jan 2025

分块正则化 5×2 交叉验证的麦克尼马尔检验：比较两种分类算法

Jing Yang、Ruibo Wang、Yijun Song、Jihong Li

摘要—在比较两种分类算法的任务中，广泛使用的麦克尼马尔检验旨在推断两种分类算法的错误率之间是否存在显著差异。然而，传统麦克尼马尔检验的检验功效通常并不理想，因为检验中的留出法（HO）仅使用单个训练-验证划分，而该划分通常会对错误率产生高度变化的估计。相比之下，交叉验证（CV）方法会多次重复留出法，并产生稳定的估计。因此，交叉验证方法在提高麦克尼马尔检验的检验功效方面具有显著优势。在所有类型的交叉验证方法中，块正则化的 5×2 交叉验证（分块交叉验证）已在许多先前研究中显示出，在算法比较任务中优于其他交叉验证方法，因为 5×2 分块交叉验证可以通过对所有训练集之间的重叠记录数量进行正则化，生成高质量的错误率估计量。在本研究中，我们压缩 5×2 分块交叉验证中的10个相关列联表，以形成一个有效列联表。然后，我们基于该有效列联表定义 5×2 分块交叉验证麦克尼马尔检验。我们通过多组模拟和真实世界数据集，证明了所提出的 5×2 分块交叉验证麦克尼马尔检验具有合理的第一类错误和颇具前景的检验功效。

索引词—麦克尼马尔检验， 5×2 分块交叉验证，算法比较，列联表，错误率，相关性。

1 引言

在N机器学习领域中，常用的麦克尼马尔检验旨在比较两种分类算法的错误率，并选出表现更好的一种。该检验已广泛应用于各种真实世界的任务，包括图像分类 [1], [2],语音与情感识别 [3], [4],情感分类 [5], [6], 以及词语切分任务[7]。

传统的麦克尼马尔检验通常采用留出法（HO）验证，即仅在数据集上进行一次训练-验证划分。训练集用于训练参与比较的两种分类算法，验证集则用于构建 2×2 列联表，以汇总这两种算法预测结果不一致的样本数量。由于算法错误率的留出法估计量通常不稳定 [8],[9], [10], ，传统麦克尼马尔检验的检验功效并不理想 [11]。

与留出法验证相比，交叉验证（CV）会多次重复留出法验证，并对所有留出法估计量取平均，从而得到更稳定的错误率估计量 [12]。因此，交叉验证方法有望显著提高麦克尼马尔检验的检验功效。本研究旨在利用交叉验证方法改进麦克尼马尔检验。然而，由于交叉验证方法中的所有留出验证都在同一个数据集上进行，由多次留出验证产生的列联表和麦克尼马尔检验统计量彼此相关。因此，构建基于交叉验证的麦克尼马尔检验时，主要挑战在于如何汇总这些相关的列联表，以构成合理的麦克尼马尔检验统计量。

一种直观的应对方法是，将交叉验证方法中的所有麦克尼马尔检验统计量相加。具体而言，对于广受欢迎的 K折交叉验证 [8], [13], ，人们会天真地假设K K折交叉验证中的麦克尼马尔检验统计量相互独立，并将其相加，构造一个服从自由度为 K 的 χ^2 分布的新检验统计量。然而，当 $K > 2$ 时，由于 K折交叉验证中的 K 训练集存在大量重叠，这些 K 麦克尼马尔检验统计量彼此相关。因此，构造出的检验统计量并不合理。本研究将这种直观的麦克尼马尔检验称为天真的 K折交叉验证麦克尼马尔检验，并不建议使用。

先前的研究表明，分块正则化的 5×2交叉验证（分块交叉验证）在算法比较 [14]任务中比 K 折交叉验证具有更多优势。具体而言，Dietterich表明，在用于比较分类算法 [9]的配对 t-检验中， 5×2 交叉验证优于 K 折交叉验证。Alpaydin 进一步采用 5×2 交叉验证开发了一种组合F-检验 [15]。此外，Wang等人开发了一种 3×2 分块交叉验证t-检验，并表明其第一类错误和检验功效与 5×2 交叉验证配对 t-检验和 F-检验具有可比性 [16]。Wang、Li和Li 基于 5×2 分块交叉验证[17], 开发了一种校准的 F-检验，并且该检验在算法比较任务中实现了最先进的性能。事实上，一个 5×2 分块交叉验证会对所有训练集之间的重叠记录数进行正则化，因此具有最小方差性质，并能为算法的错误率诱导出一个稳定的估计量[14]。因此，我们考虑使用一个 5×2 分块交叉验证来开发一个

• *J. Yang* 任职于中国太原市山西大学自动化与软件工程学院，邮编 *030031*。*R. Wang*、*Y. Song* 和 *J. Li* 任职于中国太原市山西大学现代教育技术学院，邮编 *030006*。电子邮箱： {*yangjing, wangruibo, songyj, li ml*}*@sxu.edu.cn*。稿件于*2005*年*4*月*19*日收到；修订稿于*2015*年*8*月*26*日收到。

本研究中的新型麦克尼马尔检验。

鉴于 5×2 分块交叉验证中的十个列联表彼此相关，我们并不直接将十个麦克尼马尔检验统计量相加，而是通过合理的贝叶斯分析，将这十个相关的列联表压缩成一个称为有效列联表的虚拟表；在此过程中，我们定义并分析了十个列联表中不一致概率的估计量之间的两个相关系数。有效列联表帮助我们定义了一种类似于传统麦克尼马尔检验的新型麦克尼马尔检验。在对相关系数进行合理设定的情况下，这种新型麦克尼马尔检验被称为 5×2 分块交叉验证麦克尼马尔检验。

在多个合成数据集和真实世界数据集上开展了大量实验，涉及几种常用的分类算法。使用第一类错误和功效曲线来衡量用于比较两种分类算法的显著性检验的性能。此外，实验还研究了14种现有的显著性检验。实验结果表明，所提出的 5×2 分块交叉验证麦克尼马尔检验优于其他检验方法。

本文其余部分组织如下。第2节研究传统的麦克尼马尔检验。随后，我们在第3节介绍一种朴素的 naïve K-折交叉验证麦克尼马尔检验。第4 节详细阐述我们提出的 5×2 分块交叉验证麦克尼马尔检验。第5节报告我们的实验和观察结果，随后在第 6 节给出结论。所有定理和引理的证明见附录。

2 经典麦克尼马尔检验

2.1 留出法验证中的麦克尼马尔检验

设 $D_n = \{z_i : \ z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^n$ 为一个数据集，其中包含从未知总体$\mathfrak{D}$中抽取的 n 个独立同分布记录。在一条记录 z_i 中， $\boldsymbol{x}_i$ 是预测变量向量， y_i 是 $\boldsymbol{x}_i$的真实类别标签。设 $\mathcal{A}$ 和 $\mathcal{B}$ 为在 D_n上进行比较的两个分类算法。比较任务可形式化为以下假设检验问题:

$$H_0 : \ \mu_{\mathcal{A}} - \mu_{\mathcal{B}} = 0 \ \ vs. \ H_1 : \ \mu_{\mathcal{A}} - \mu_{\mathcal{B}} \neq 0, \qquad (1)$$

其中， $\mu_{\mathcal{A}}$ 和 $\mu_{\mathcal{B}}$ 分别是 $\mathcal{A}$ 和 $\mathcal{B}$的错误率，也称为 $\mathcal{A}$和 $\mathcal{B}$ 在零一损失 [18]下的泛化误差。

式（1）中的问题可以通过麦克尼马尔检验来解决。经典麦克尼马尔检验与留出法验证结合使用。在留出法验证中，D_n 被划分为两个子块，划分方式为 (S, T)，其中S 和 T 分别对应训练集和验证集，且 $S \cup T = D_n$ 和 $S \cap T = \emptyset$。训练集和验证集的大小分别为 $n_1 = |S|$ 和 $n_2 = |T|$。

基于划分 (S, T)，算法 $\mathcal{A}$ and$\mathcal{B}$ 在训练集 S 上进行训练，并生成两个分类器，即 $\mathcal{A}(S)$ 和 $\mathcal{B}(S)$。随后，在验证集 T上评估这两个分类器。验证结果可汇总为列联表$\mathcal{C} = (n_{00}, n_{01}, n_{10}, n_{11})$，使得 $n_{00}+n_{01}+n_{10}+n_{11} = n_2$。 $\mathcal{C}$ 的解释见表 1。

表 1 列联表 $\mathcal{C}$的解读

		模型 $\mathcal{B}(S)$	
		不正确	正确
Model $\mathcal{A}(S)$	Incorrect	n_{00}: 中的记录数量 T 被两者错误分类 $\mathcal{A}$ 和 $\mathcal{B}$.	n_{01}: 记录数 在 T 中被 $\mathcal{A}$错误分类 但不是通过 $\mathcal{B}$。
	Correct	n_{10}: 其中的记录数 T被 $\mathcal{B}$ 错误分类但 不是由 $\mathcal{A}$。	n_{11}: 其中的记录数 T 被以下两者均误分类: $\mathcal{A}$和 $\mathcal{B}$。

基于 $\mathcal{C}$，传统的麦克尼马尔检验使用以下统计量 [19]:

$$\mathcal{M}^{\text{HO}} = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \sim \chi^2(1), \qquad (2)$$

其中，分子中的 -1 是连续性校正 [20]。如果$\mathcal{M}^{\text{HO}} > \chi^2_\alpha$ (1) 其中 α 是显著性水平，则拒绝 H_0 。

2.2 $\mathcal{C}$的贝叶斯解释

对于记录 z，令 y 为真实类别标签，令$\hat{}y_{\mathcal{A}}$ 和$\hat{}y_{\mathcal{B}}$ 分别为分类器 $\mathcal{A}(S)$ 和 $\mathcal{B}(S)$的预测类别标签。那么，列联表 $\mathcal{C}$ 可视为从参数为 $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ 的多项分布中抽取的一个样本，其中 $\pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1$。形式上，我们记为 $\mathcal{C}|\boldsymbol{\pi} \sim \mathbf{M}(n_2, \boldsymbol{\pi})$，或者

$$P(\mathcal{C}|\boldsymbol{\pi}) = \frac{n_2!}{n_{00}!n_{01}!n_{10}!n_{11}!}\pi_{00}^{n_{00}}\pi_{01}^{n_{01}}\pi_{10}^{n_{10}}\pi_{11}^{n_{11}}, \qquad (3)$$

其中 $\boldsymbol{\pi}$ 定义如下:

$$\begin{aligned}
\pi_{00} &= P(\hat{y}_{\mathcal{A}} \neq y, \hat{y}_{\mathcal{B}} \neq y), \\
\pi_{01} &= P(\hat{y}_{\mathcal{A}} \neq y, \hat{y}_{\mathcal{B}} = y), \\
\pi_{10} &= P(\hat{y}_{\mathcal{A}} = y, \hat{y}_{\mathcal{B}} \neq y), \\
\pi_{11} &= P(\hat{y}_{\mathcal{A}} = y, \hat{y}_{\mathcal{B}} = y).
\end{aligned}$$

由 $\boldsymbol{\pi}$可导出若干有意义的随机变量，包括一个不一致概率 e 和三个条件错误率 r、 q_a和 q_b。相应地，这些随机变量的估计量可由 $\mathcal{C}$导出，即 E、 R、 Q_a和 Q_b。表 2 展示了这些随机变量和估计量的解释。由于$\mu_{\mathcal{A}} = \pi_{00} + \pi_{01}$ 和 $\mu_{\mathcal{B}} = \pi_{00} + \pi_{10}$，式 (1) 中的问题可改写为

$$H_0 : \ r = 0.5 \ \ vs. \ H_1 : \ r \neq 0.5. \qquad (4)$$

在估计量 E、 R、 Q_a和 Q_b 中， $n_{01} + n_{10}$、 n_{00} 和 n_{01} 的条件分布如下。

引理*1*。 鉴于 $\mathcal{C}|\boldsymbol{\pi} \sim \mathbf{M}(n_2, \boldsymbol{\pi})$，我们得到

$$\begin{aligned}
n_{01} + n_{10} &\sim \mathbf{B}(n_2, e), \\
n_{01}|n_{01} + n_{10} &\sim \mathbf{B}(n_{01} + n_{10}, r), \\
n_{01}|n_{01} + n_{11} &\sim \mathbf{B}(n_{01} + n_{11}, q_a), \\
n_{00}|n_{10} + n_{00} &\sim \mathbf{B}(n_{01} + n_{00}, q_b),
\end{aligned}$$

其中 $\mathbf{B}(\cdot, \cdot)$ 表示二项分布。

从贝叶斯视角来看， e、 r、 q_a 和 q_b的先验分布被假设为共轭贝塔分布 $Be(\lambda, \lambda)$。

表 2 定义在 $\boldsymbol{\pi}$ 上的几个有意义的随机变量及其估计量的解释。

符号表示	定义	概率形式	估计量	解释
e	$\pi_{01}+\pi_{10}$	$p(\hat{y}_{\mathcal{A}}\neq\hat{y}_{\mathcal{B}})$	$E=\frac{n_{01}+n_{10}}{n_2}$	算法 $\mathcal{A}$ 和 $\mathcal{B}$之间预测结果的不一致概率
r	$\frac{\pi_{01}}{\pi_{01}+\pi_{10}}$	$p(\hat{y}_{\mathcal{A}}\neq y\|\hat{y}_{\mathcal{A}}\neq\hat{y}_{\mathcal{B}})$	$R=\frac{n_{01}}{n_{01}+n_{10}}$	算法 $\mathcal{A}$ 在不一致预测结果上的条件错误率
q_a	$\frac{\pi_{01}}{\pi_{01}+\pi_{11}}$	$p(\hat{y}_{\mathcal{A}}\neq y\|\hat{y}_{\mathcal{B}}=y)$	$Q_a=\frac{n_{01}}{n_{01}+n_{11}}$	算法 $\mathcal{A}$ 在算法 $\mathcal{B}$ 正确时的条件错误率
q_b	$\frac{\pi_{00}}{\pi_{10}+\pi_{00}}$	$p(\hat{y}_{\mathcal{A}}\neq y\|\hat{y}_{\mathcal{B}}\neq y)$	$Q_b=\frac{n_{00}}{n_{10}+n_{00}}$	算法 $\mathcal{A}$ 在算法 $\mathcal{B}$ 错误时的条件错误率

因此， e、 r q_a和 q_b在给定 $\mathcal{C}$ 条件下的后验分布见引理 2。

引理 **2.** 在贝塔先验 $Be(\lambda,\lambda)$下， e、 r、 q_a和 q_b 的后验分布为

$$\begin{aligned} e|\mathcal{C} &\sim Be(n_{01}+n_{10}+\lambda, n_{00}+n_{11}+\lambda),\\ r|\mathcal{C} &\sim Be(n_{01}+\lambda, n_{10}+\lambda),\\ q_a|\mathcal{C} &\sim Be(n_{01}+\lambda, n_{11}+\lambda),\\ q_b|\mathcal{C} &\sim Be(n_{00}+\lambda, n_{10}+\lambda). \end{aligned}$$

推论 **1.** 在参数为 λ =1 的非信息性贝塔先验下， e、 r、 q_a 和 q_b 的后验众数为

$$\begin{aligned} mode[e|\mathcal{C}] &= E,\\ mode[r|\mathcal{C}] &= R,\\ mode[q_a|\mathcal{C}] &= Q_a,\\ mode[q_b|\mathcal{C}] &= Q_b. \end{aligned}$$

需要指出的是，列联表的贝叶斯解释并非本研究的主要贡献，因为已有研究开发了对混淆矩阵进行类似贝叶斯解释的方法，用于分析精确率、召回率和 F_1 分数[21] [22] [23]。

3 朴素 K 折交叉验证麦克尼马尔检验

本节介绍一种将麦克尼马尔检验与交叉验证方法相结合的直观方式。以广受欢迎的 K 折交叉验证为例，它会生成 K 个列联表，记为 $\mathcal{C}_k = (n_{00,k},n_{01,k},n_{10,k},n_{11,k})$ ，其中包含$k=1,\ldots,K$。此外，根据公式（2）， K麦克尼马尔检验统计量可定义为 $\mathcal{M}_1^{\mathrm{HO}},\ldots,\mathcal{M}_K^{\mathrm{HO}}$。进而，我们将 K 个统计量相加，构造如下统计量。

$$\mathcal{M}^{\mathrm{KCV}}=\sum_{k=1}^{K}\mathcal{M}_k^{\mathrm{HO}}=\sum_{k=1}^{K}\frac{(|n_{01,k}-n_{10,k}|-1)^2}{n_{01,k}+n_{10,k}}. \quad (5)$$

若 $\mathcal{M}_1^{\mathrm{HO}},\ldots,\mathcal{M}_K^{\mathrm{HO}}$ 相互独立，则 $\mathcal{M}^{\mathrm{KCV}}\sim\chi^2(K)$。因此，若 $\mathcal{M}^{\mathrm{KCV}}>\chi^2_\alpha(K)$，则 H_0 被拒绝。

上述统计量是朴素的，因为在K折交叉验证中，训练集之间存在大量重叠，导致 $\mathcal{M}_1,\ldots,\mathcal{M}_K$彼此相关。因此， $\mathcal{M}^{\mathrm{KCV}}$ 并非精确的 $\chi^2(K)$。本研究将针对 $\mathcal{M}^{\mathrm{KCV}}$ 的检验称为朴素的¨ K折交叉验证麦克尼马尔检验，并不建议使用。

表 3 分块交叉验证（BCV）的划分集 $\mathbb{P}$。

j	第 S 折	第 T 折
1	D_1, D_2, D_3, D_4	D_5, D_6, D_7, D_8
2	D_1, D_3, D_5, D_7	D_2, D_4, D_6, D_8
3	D_1, D_2, D_5, D_6	D_3, D_4, D_7, D_8
4	D_1, D_4, D_5, D_8	D_2, D_3, D_6, D_7
5	D_1, D_3, D_6, D_8	D_2, D_4, D_5, D_7

4 5×2 分块交叉验证 麦克尼马尔检验

4.1 5×2 分块交叉验证

以往研究建议在比较两种算法时采用随机划分的 5×2 交叉验证（RCV） [9][15] [24]。然而， 5×2 RCV 的随机划分可能导致训练集之间过度重叠，从而降低算法比较方法的性能。相比之下， 5×2 分块交叉验证（BCV）会使任意两个训练集之间重叠记录的数量保持一致，因此， 5×2BCV 的错误率估计量方差小于 5×2 RCV 估计量的方差[14] [17]。因此， 5×2 BCV 为开发新型算法比较方法提供了一种颇具前景的数据划分方案。

形式上， 5×2 分块交叉验证会进行五次重复，每次都采用包含五个正则化划分的两折交叉验证划分。令$\mathbb{P}=\{(S_j,T_j)\}_{j=1}^{5}$ 表示满足 $|S_j|=|T_j|=n/2$ 和$|S_j\cap S_{j'}|\approx n/4$ 正则化条件的 5×2 分块交叉验证划分集，其中 $j\neq j'$。每个划分 (S_j,T_j) 对应一次两折交叉验证，其中 S_j 和 T_j 以轮转方式用作训练集。

为了构建 $\mathbb{P}$，首先将 D_n 划分为八个大小相等的子块 D_1，……， D_8。根据表 3 中的启发式规则，将这些子块组合形成 $\mathbb{P}$。这些规则取自试验设计（DoE）领域中二水平正交表$L_8(2^7)$的前五列 [25]。

4.2 在 5×2 分块交叉验证上的列联表

在 5×2 分块交叉验证中比较算法 $\mathcal{A}$ 和 $\mathcal{B}$ 时，会得到一组由十个列联表组成的集合，即 $\boldsymbol{\mathcal{C}}_{5\times2}=\{\mathcal{C}_k^{(j)}\}_{j=1}^{5}=\{\mathcal{C}_1^{(j)},\mathcal{C}_2^{(j)}\}_{j=1}^{5}$ $\mathcal{C}_k^{(j)}=\boldsymbol{\mathcal{C}}(,)$， $\left(n_{00,k}^{(j)},n_{01,k}^{(j)},n_{10,k}^{(j)},n_{11,k}^{(j)}\right)$ $k=1\ 2$ $\mathcal{C}_1^{(j)}$ S_j 其中 ,。表格使用 和 T_j作为训练集和验证集。表格 $\mathcal{C}_2^{(j)}$使用 T_j和 S_j 来训练和验证算法。

¯此外，根据$(\bar{n}_{00},\ \bar{n}_{01},\ \bar{n}_{10},\ \bar{n}_{11})$ 可得到一个$\boldsymbol{\mathcal{C}}_{5\times2}$平均列联表，而它满足

isfies $\bar{n}_{00}+\bar{n}_{01}+\bar{n}_{10}+\bar{n}_{11}=n/2$and

$$\bar{n}_{ii'}=\frac{1}{10}\sum_{j=1}^{5}\sum_{k=1}^{2}n_{ii',k}^{(j)},\quad \forall i,i'=0,1. \tag{6}$$

¯相应地， $\mathcal{C}_{5\times2}$中的 e、 r、 q_a和 q_b 的估计量定义如下。

$$\begin{aligned}
E_{5\times2} &= \frac{2(\bar{n}_{01}+\bar{n}_{10})}{n},\\
R_{5\times2} &= \frac{\bar{n}_{01}}{\bar{n}_{01}+\bar{n}_{10}},\\
Q_{a,5\times2} &= \frac{\bar{n}_{01}}{\bar{n}_{01}+\bar{n}_{11}},\\
Q_{b,5\times2} &= \frac{\bar{n}_{00}}{\bar{n}_{10}+\bar{n}_{00}}.
\end{aligned}$$

$E_{5\times2}$ 的分母为常数，因此

$$E_{5\times2}=\sum_{j=1}^{5}\sum_{k=1}^{2}\frac{(n_{01,k}^{(j)}+n_{10,k}^{(j)})}{5n}=\sum_{j=1}^{5}\sum_{k=1}^{2}\frac{E_k^{(j)}}{10}, \tag{7}$$

其中， $E_k^{(j)}$ 是 $\mathcal{C}_k^{(j)}$ 上 e 的估计量。

定理*1*。 $E_{5\times2}$ 的方差表示如下。

$$\mathrm{Var}[E_{5\times2}|e]=\frac{1+\rho_1+8\rho_2}{5n}e(1-e), \tag{8}$$

其中， ρ_1 和 ρ_2 的定义如下。

- Let $\sigma^2=\mathrm{Var}\left[E_k^{(j)}\right]=2e(1-e)/n$ be the variance of an HO estimator of e;
- $\rho_1=\mathrm{Cov}\left[E_1^{(j)},E_2^{(j)}\right]/\sigma^2$ is the inter-group correlation coefficient of two HO estimators in $E^{(j)}$ in a two-fold CV.
- $\rho_2=\mathrm{Cov}\left[E_k^{(j)},E_{k'}^{(j')}\right]/\sigma^2$ is the intra-group correlation coefficient of two HO estimators of e in different two-fold CVs where $j\neq j'$ and $k,k'=1,2$.

式（8）表明， ρ_2 比 ρ_1发挥着更重要的作用。此外，由于 ρ_1 和 ρ_2 与特定的算法类型和数据集分布有关，因此很难推导出 ρ_1 和 ρ_2的闭式表达式。尽管如此，我们仍能推导出 ρ_1 和 ρ_2的一些有用界限，而这些界限在两种算法比较中的统计推断程序中不可或缺。

4.3 ρ_1 和 ρ_2的性质

鉴于 5×2 分块交叉验证中的五个划分是在同一个数据集上进行的，$\boldsymbol{\mathcal{C}}_{5\times2}$上的所有 $E_k^{(j)}$s彼此相关。因此， ρ_1和 ρ_2 通常偏离零值。

以下阐述了 ρ_1 和 ρ_2 的若干理论性质。

引理*3*。假设 $E_k^{(j)}$ 仅依赖于验证集和 n，则 $\rho_1=0$和 $\rho_2=1/2$。

然而， 5×2 分块交叉验证中训练集之间的相关性也会受到训练集的影响，因为这些训练集存在大量重叠。此外，在一个温和的假设下，定理2中的界成立。

定理*2*。假设 $E_k^{(j)}$ 取决于数据集大小 n、验证集以及重叠记录的数量

在一次 5×2 分块交叉验证中，训练集之间的相关系数 ρ_1 和 ρ_2 满足以下界。

$$\rho_2<(1+\rho_1)/2. \tag{9}$$

此外，如果 $\rho_1=0$，那么 $\rho_2<1/2$。

许多先前的研究在广泛的损失函数、不同的算法和各种数据集上考察了 ρ_1 和 ρ_2 。这些研究证实了定理2中的界成立，并指出该界较为宽松。此外，这些研究将该界略微收紧为以下形式。

$$\rho_1\leq1/2\ \text{ and }\ 0\leq\rho_2\leq1/2. \tag{10}$$

他们进一步发现，这些紧界以很高的概率成立。因此，他们建议在实际比较两种算法时采用公式（10）中的界。图 1 也验证了 ρ_1 和 ρ_2 的大多数取值都满足公式（10）中的界。

4.4 5×2 BCV 上的有效列联表

本节介绍“有效列联表”的概念，用于近似在 $\boldsymbol{\mathcal{C}}_{5\times2}$ 中的十个相关列联表条件下变量 e、 r、 q_a 和 q_b 的后验分布。有效列联表是一个单一的虚拟列联表，记为 $\mathcal{C}_e=(n_{00,\ e},\ n_{01,\ e},\ n_{10,\ e},\ n_{11,\ e})$ 。此外，令 $n_e=n_{00,\ e}+n_{01,\ e}+n_{10,\ e}+n_{11,\ e}$ 表示 $\mathcal{C}_e$ 的有效验证集大小。

$\mathcal{C}_e$ 上 e、 r、 q_a 和 q_b 的估计量定义为 E_e、R_e、 $Q_{a,\ e}$ 和 $Q_{b,\ e}$ ，其形式如下。

$$\begin{aligned}
E_e &= \frac{n_{01,e}+n_{10,e}}{n_e},\\
R_e &= \frac{n_{01,e}}{n_{01,e}+n_{10,e}},\\
Q_{a,e} &= \frac{n_{01,e}}{n_{01,e}+n_{11,e}},\\
Q_{b,e} &= \frac{n_{00,e}}{n_{10,e}+n_{00,e}}.
\end{aligned}$$

在关于 $\mathcal{C}_e$ 条件下 e、r、 q_a 和 q_b 的众数以及估计量 E_e 的方差的某些约束下， $\mathcal{C}_e$ 的闭式表达式如下所示。

定理 *3*。 采用非信息性贝塔先验，并在以下约束条件下

$$\begin{aligned}
mode[e|\mathcal{C}_e] &= E_{5\times2},\\
mode[r|\mathcal{C}_e] &= R_{5\times2},\\
mode[q_a|\mathcal{C}_e] &= Q_{a,5\times2},\\
mode[q_b|\mathcal{C}_e] &= Q_{b,5\times2},\\
Var[E_e|e] &= Var[E_{5\times2}|e],
\end{aligned}$$

$\mathcal{C}_e$ 的闭式表达式是

$$\mathcal{C}_e=\frac{10}{1+\rho_1+8\rho_2}\bar{\mathcal{C}}_{5\times2}, \tag{11}$$

其中$n_{00,\ e}$, $n_{01,\ e}$, $n_{10,\ e}$， 以及$n_{11,\ e}$ 是

$$n_{ii',e}=\frac{10\bar{n}_{ii'}}{1+\rho_1+8\rho_2},\quad \forall i,i'=0,1.$$

根据式 (11)，可得有效验证集大小为

$$n_e = \frac{5n}{1+\rho_1+8\rho_2}. \tag{12}$$

4.5 5×2 分块交叉验证麦克尼马尔检验

在给定 $\mathcal{C}_e$ 和 ρ_1 和 ρ_2的条件下，类似于公式（2）中的 $\mathcal{M}^{\text{HO}}$，定义如下卡方统计量。

$$\mathcal{M}^{\text{BCV}}_{\rho_1,\rho_2} = \frac{(|n_{01,e}-n_{10,e}|-1)^2}{n_{01,e}+n_{10,e}} \sim \chi^2(1), \tag{13}$$

将式 (11) 代入式 (13) 可得到 $\mathcal{M}^{\text{BCV}}_{\rho_1,\rho_2}$ 的等价形式。

$$\mathcal{M}^{\text{BCV}}_{t} = \frac{t(|\bar{n}_{01}-\bar{n}_{10}|-1/t)^2}{\bar{n}_{01}+\bar{n}_{10}} \sim \chi^2(1), \tag{14}$$

其中 $t=10/（1+\rho_1+8\rho_2）$。

然而， $\mathcal{M}^{\text{BCV}}_{t}$包含一个未知参数 t，因此不能用于显著性检验。随着 t的增加，统计量 $\mathcal{M}^{\text{BCV}}_{t}$变大，由 $\mathcal{M}^{\text{BCV}}_{t}$导出的麦克尼马尔检验趋向于成为一种宽松检验，容易产生假阳性结论。相比之下，根据算法比较任务中的“保守原则”，[18], 更偏好 t 的较小值。因为t 是关于 ρ_1 和ρ_2的单调递减函数。考虑公式（10）中的界限，我们设定 $\rho_1=\rho_2=0.5$, 因此 $t=20/11$。得到以下检验统计量。

$$\mathcal{M}^{\text{BCV}} = \frac{20(|\bar{n}_{01}-\bar{n}_{10}|-11/20)^2}{11(\bar{n}_{01}+\bar{n}_{10})} \sim \chi^2(1), \tag{15}$$

在本研究中， $\mathcal{M}^{\text{BCV}}$ 被称为 5×2 分块交叉验证 McNemar检验统计量。此外， 5×2 分块交叉验证 McNemar检验的判定规则为：若 $\mathcal{M}^{\text{BCV}} > \chi^2_\alpha(1)$，则拒绝 H_0。此外，使用 $\alpha=0.05$ 和 $\chi^2_{0.05}(1) \approx 3.841$。

5 实验结果与分析

为验证所提出的 5×2 分块交叉验证麦克内马尔检验的优越性，我们提出了两个研究问题（RQs）。

RQ1：方程（8）中的相关系数 ρ_1 和 ρ_2 如何分布?

RQ2： 5×2 分块交叉验证麦克尼马尔检验是否优于现有的两种分类算法比较方法?

用于比较两种分类算法的最先进方法主要包含4类共17种不同的显著性检验。现有检验及其推荐设置如下。

(I) t**检验族**（9 项检验）。

(1) 重复 HO（RHO）配对 t-检验，采用 15 次重复 HO，并使用 $n_1=2n/3$ [9]。(2) K折交叉验证（CV）配对 t-检验，使用 $K=10$ [9]。(3) 5×2 交叉验证（CV）配对 t-检验 [9]。(4) 校正 RHO t-检验，使用 $n_1=9n/10$ 和 15 次重复 HO [18]。(5) 伪自助法检验 [18]。(6) 校正伪自助法检验 [18]。(7) 校正 10×10 交叉验证（CV） t-检验 [26]。(8) 组合 5×2 交叉验证（CV） t-检验 [24]。

(9) 分块 3×2 交叉验证 t-检验 [16]。

(II) F**-检验系列** (2项检验)

(1) 组合 5×2 交叉验证 F-检验 [15]。

(2) 校准 5×2 分块交叉验证 F-检验 [17]。

（III） Z**检验系列**（3 项检验）

(1) 使用 $n_1=2n/3$ [9]的比例检验。(2)保守 Z-检验 [18]。(3) K-折交叉验证-CI Z-检验，使用 $K=10$ 和“out”类型的方差估计量 [27]。

(IV) **McNemar 检验族**(3 tests)

（1）使用 $n_1=2n/3$[9] 的常规 HO麦克尼马尔检验。（2）使用 $K=10$的 K折交叉验证朴素¨ 麦克尼马尔检验。（3） 5×2 分块交叉验证麦克尼马尔检验。

由于Nadeau 和Bengio发现，这三种检验的检验功效低于校正的RHO t-检验[18]且成本更高，因此我们在所有实验中排除了保守的 Z-检验、伪自助法检验和校正伪自助法检验。

5.1 RQ1 实验

为展示 ρ_1 和 ρ_2的分布，使用了29个真实世界UCI数据集和七种常用的分类算法。所有真实世界 UCI 数据集均列于表4中。对于每个数据集，省略包含缺失值的记录。每个数据集中的记录数、预测变量数和类别数列于表4中。实验中使用的七种分类算法详述如下。

(1) **多数类分类器**。测试记录被分配到训练集中频率最高的类别标签。该分配规则与输入预测变量无关。

(2) **均值分类器**。针对每个类别，计算其预测变量的均值向量。将测试记录分配给其均值向量与该实例之间欧几里得距离最小的类别。它可视为线性判别分析的一种特殊情况，其协方差矩阵为共享矩阵，且对角线元素相等、非对角线元素为零。

(3) **逻辑回归**。使用“RWeka”包中的逻辑回归分类器。(4) **支持向量机**。考虑采用序列最小优化算法的支持向量机。使用“RWeka”包中采用默认设置的“SMO”分类器。(5) **RIPPER**。使用 “RWeka” 包中采用默认设置的 “JRip” 分类器。(6)**C4.5**。使用“RWeka”包中采用默认设置的“J48”分类器。(7) **K近邻**。使用“kknn”包中的k近邻分类器。参数 k 设为 5，并采用三角核函数。

在每个数据集上，对任意两个算法进行比较，因此进行了 $\binom{7}{2}=21$ 种不同的算法成对比较.总共进行了 $21\times29=609$ 次实验。在每次实验中，我们从一个数据集中无放回地随机抽取300条记录，然后对这些记录执行块-正则化划分并计算 es-

表4 投稿论文实验中使用的数据集。

No.	Name	#记录	#预测变量	#类别
1	artificial	30654	7	10
2	australian	690	14	2
3	balance	625	4	3
4	car	1728	6	4
5	cmc	1473	9	3
6	信用	690	15	2
7	捐赠者	748	4	2
8	耀斑	1388	10	2
9	德国	1000	20	2
10	波兰字母识别	3196	36	2
11	字母	20000	16	26
12	魔术	19020	10	2
13	乳腺摄影	830	5	2
14	托儿所	12960	8	5
15	光学手写数字	5620	64	10
16	页面块	5473	10	5
17	手写数字	10992	16	10
18	pima	768	8	2
19	环形正态	7400	20	2
20	卫星47	4435	36	6
21	垃圾邮件数据集	4601	57	2
22	剪接	3190	60	3
23	井字棋	958	9	2
24	泰坦尼克号	2201	3	2
25	输血	748	4	2
26	双正态	7400	20	2
27	wave	5000	40	3
28	葡萄酒质量	4898	11	7
29	酵母	1484	7	10

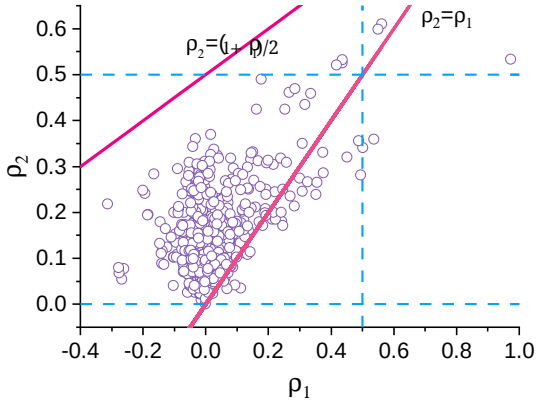

图1. ρ_1 和 ρ_2的散点图。

timations $E_i^{(j)}$。该过程重复1,000次。在这1,000次重复中获得 ρ_1 和 ρ_2 的值。

图 1 展示了 609 个实验中ρ_1 和 ρ_2 数值的分布。得到以下三点观察结果。

（1） ρ_1 和 ρ_2 的大多数值都分布在 0.5 以下。此外， ρ_1 在 0 附近变化，而 ρ_2 大于 0。因此，公式（10）中的界以高概率成立。该观察结果与先前的研究一致。（2） ρ_1 和 ρ_2 的所有值都位于红色实线 $\rho_2 = (1+\rho_1)/2$右下方。该观察结果证实了定理2中的界。此外，实线与散点之间有较大的间距，这表明定理2中的界是宽松。（3） ρ_1 和 ρ_2 的多数值满足 $\rho_2 \geq \rho_1$。

5.2 RQ2 的实验

使用三个合成数据集和两个真实世界数据集来回答 RQ2。

合成数据集 1：Epsilon 数据集。Epsilon 数据集由 Dietterich [9] 开发，用于研究比较两种分类算法的假设检验的第一类错误。对于 Epsilon 数据集，基于数据集大小 n，我们直接生成两种算法的零一损失向量。具体而言，令 i 为数据记录索引，其中 $i = 1, \ldots, n$。令 $\mathbf{1}_{\mathcal{A}}(i)$ 和 $\mathbf{1}_{\mathcal{B}}(i)$ 分别为算法 $\mathcal{A}$ 和 $\mathcal{B}$ 在第 i 条记录上的损失值。然后，根据以下规则生成损失值。

- When $1 \leq i \leq n/2$, $\mathbf{1}_{\mathcal{A}}(i) \sim \mathbf{B}(1, \epsilon/2)$ and $\mathbf{1}_{\mathcal{B}}(i) \sim \mathbf{B}(1, 3\epsilon/2)$.
- When $n/2 + 1 \leq i \leq n$, $\mathbf{1}_{\mathcal{A}}(i) \sim \mathbf{B}(1, 3\epsilon/2)$ and $\mathbf{1}_{\mathcal{B}}(i) \sim \mathbf{B}(1, \epsilon/2)$.

根据规则，我们可以得出：在数据集的前半部分，算法 $\mathcal{A}$和 $\mathcal{B}$的错误率分别为 $\mu_{\mathcal{A}} = \epsilon/2$ 和 $\mu_{\mathcal{A}} = 3\epsilon/2$；在数据集的剩余部分， $\mu_{\mathcal{A}} = 3\epsilon/2$ 和 $\mu_{\mathcal{A}} = \epsilon/2$；在整个数据集上，我们得到 $\mu_{\mathcal{A}} = \mu_{\mathcal{B}} = \epsilon$，因此，式(1)中的原假设$H_0$ 成立。因此，epsilon 数据集只能用于获得检验的第一类错误。在此数据集中，使用了 $n = 300$和 $\epsilon = 0.1$ 的设置。

合成数据集 2：EXP6 数据集。EXP6 数据集是一个包含两个连续预测变量的六分类学习问题。它由 [28] 开发，并用于 [9]中的算法比较任务。具体而言，在一个 EXP6 数据集 $D_n = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$中，令 $\boldsymbol{x} = (x_1, x_2)$ 为两个预测变量，y 为满足 $y \in \{Y_j\}_{j=1}^6$的真实类别标签。EXP6 数据集中的预测变量x_1 和 x_2 在区域$\mathbf{X} = [0, 15] \times [0, 15]$上以分辨率 0.1 的均匀网格空间中均匀分布。类别标签 $y_{\mathbf{x}}$ 按如下方式生成。

$$y_{\mathbf{x}} = \begin{cases} Y_1 & x_2 - f_1(x_1) \geq 0 \ \wedge \ x_2 - f_2(x_1) \geq 0 \\ Y_2 & x_2 - f_1(x_1) < 0 \ \wedge \ x_2 - f_2(x_1) \geq 0 \\ & \quad \wedge \ x_2 - f_3(x_1) \geq 0 \\ Y_3 & x_2 - f_1(x_1) \geq 0 \ \wedge \ x_2 - f_2(x_1) < 0 \\ Y_4 & x_2 - f_1(x_1) < 0 \ \wedge \ x_2 - f_2(x_1) < 0 \\ & \quad \wedge \ x_2 - f_3(x_1) \geq 0 \\ Y_5 & x_2 - f_2(x_1) \geq 0 \ \wedge \ x_2 - f_3(x_1) < 0 \\ Y_6 & x_2 - f_2(x_1) < 0 \ \wedge \ x_2 - f_3(x_1) < 0 \end{cases} \tag{16}$$

其中 f_1、 f_2 和 f_3 是以下决策边界函数。

$$\begin{aligned} f_1(x) &= x^2 - 4x + 6, \\ f_2(x) &= 4sin(x/2) + 8, \\ f_3(x) &= -\frac{1}{25}(x^2 - 108x + 236). \end{aligned} \tag{17}$$

图 2 显示了一个带有 $n = 300$ 的 EXP6 数据集散点图。此外，我们还比较了两个算法在 EXP6 数据集上的表现。

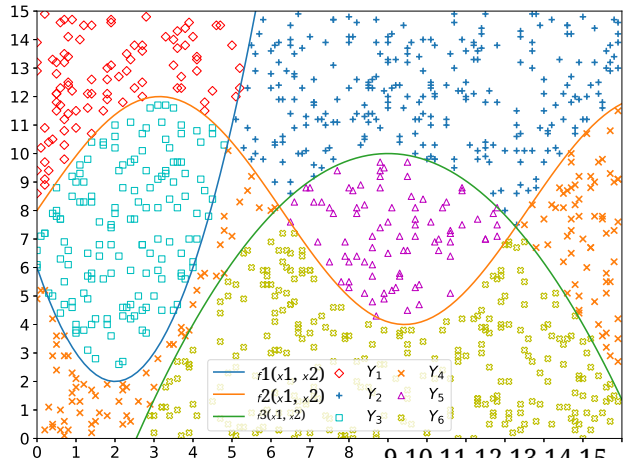

图 2. 带有 $n = 300$的 EXP6 数据集示例。

(*A*) C4.5。这是一种流行的基于树的分类算法。RWeka 包中的 J48 实现了 C4.5 算法，且未使用剪枝过程。

(*B*) FNN。第一近邻使用加权欧几里得距离，将其性能调节至与 C4.5 相当的水平。具体而言，点 $\boldsymbol{x}_i = (x_{i1}, x_{i2})$与 $\boldsymbol{x}_j = (x_{j1}, x_{j2})$ 之间的距离是

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \omega(x_{i1} - x_{j1})^2 + \omega^{-1}(x_{i2} - x_{j2})^2, \quad (18)$$

其中，$\omega \in (0.0,1.0]$ 是一个可调权重。当 $\omega = 1.0$ 时，$d(\cdot,\cdot)$ 退化为典型的欧几里得距离。

C4.5 和 FNN的错误率真实值，即 $\mu_{C4.5}$ 和$\mu_{FNN,\omega}$，是通过对在 1,000个独立同分布的 EXP6数据集上训练的分类器的所有错误率估计值取平均，并使用包含 $(15/0.1 + 1)^2 = 22{,}801$ 条记录的大型EXP6数据集进行评估而获得的。由于 FNN 的错误率取决于超参数 ω，我们为 FNN算法引入多个目标错误率。具体而言，目标错误率μ_{target} 表示为$\mu_{C4.5}$ 和$\mu_{FNN,1.0}$之间的线性插值，即 $\mu_{target} = \mu_{C4.5} - \lambda(\mu_{C4.5} - \mu_{FNN,1.0})$。我们调整 ω，使 FNN 的错误率真实值达到 μ_{target}。λ 的离散值取自 [0.0,1.0]的范围，步长为 0.1。特别地，当$\lambda = 0.0$时，FNN的目标错误率降至 $\mu_{C4.5}$，这使得式（1）中的原假设 H_0 为真。

图 3 展示了在各种目标错误率下 C4.5 和 FNN 的真实错误率。从图 3 中可以得到一些具体观察结果。C4.5 的真实错误率为 12.27%。随着 ω 增加，FNN 的真实错误率降低。当 $\omega = 1.0$ 时，FNN 的错误率为 8.21%。当 $\omega = 0.29$ 时，FNN 与 C4.5 达到相同的错误率。因此，在将 $\omega = 0.29$ 设置为给定值时，可以得到显著性检验方法的第一类错误。通过将 ω 从 0.29 增加到 1.0，可以绘制检验的功效曲线。

合成数据集3：简单数据集。令 $D_n = \{(x_i, y_i)\}_{i=1}^n$ 为一个简单的二分类数据集，其中$y_i \in \{Y_0, Y_1\}$ 为真实类别标签，x_i 为预测变量。类别标签索引 $i \sim \mathbf{B}(1,0.5)$ 是独立抽取的。于是，$x_i|y_i = Y_1 \sim N(0,1)$和 $x_i|y_i = Y_2 \sim N(\Delta,1)$，其中 $\Delta \in [0.0, 1.0]$ 为一个可调参数。此外，在简单数据集上比较两种算法。

(*A*) 逻辑回归。逻辑回归通过 R 软件中的“glm”程序包实现，并使用二项分布链接函数。

(*B*) MAJ. 多数类分类器使用训练集中出现频率最高的类别标签作为预测结果，且该预测结果与输入预测变量无关。

图 3 展示了简单数据集上逻辑回归和 MAJ 的真实错误率。无论 Δ 如何，MAJ 的错误率都保持不变，因为它仅取决于类别标签的先验分布。此外，当 Δ =0.0 时，两个类别中预测变量 x 的分布完全重叠，因此逻辑回归退化为随机猜测，其错误率与 MAJ 相同。当 Δ 增加时，逻辑回归的错误率单调下降。因此，使用 $\Delta = 0.0$ 获取检验的第一类错误，并通过将 Δ 从 0.0 增加到 1.0 来绘制功效曲线。在简单数据集中，使用 $n = 1{,}000$ 。

真实世界数据集 1：UCI 字母数据集。这是算法比较任务中一个广受欢迎的数据集 [9], [18]。它包含 20,000 个数据点、26 个类别和 16 个预测变量。以下比较两种分类算法。

(*A*) TREE。分类树使用 R 软件中的“tree”软件包，并采用默认设置。

(*B*) FNN。第一近邻使用如下扭曲距离 $d(\cdot, \cdot)$ [18]。

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{3} \upsilon^{2-k} \sum_{l \in C_k} (x_{il} - x_{jl})^2, \quad (19)$$

其中，$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 是两个预测变量向量，$C_1 = \{1,3,9,16\}$、$C_2 = \{2,4,6,7,8,10,12,14,15\}$和 $C_3 = \{5,11,13\}$。$\upsilon \in [1,50]$ 是一个可调参数，用于调整 FNN 的错误率。

此外，从字母数据集中有放回地独立同分布抽取 1,000 个包含 $n = 300$ 的数据集。通过在这 1,000 个数据集上训练算法，并在整个字母数据集上进行验证，得到 TREE 和 FNN 的错误率真实值。与 EXP6 数据集类似，真实错误率的模拟基于若干由插值参数 λ 表征的目标错误率。真实错误率如图 3 所示，图中显示 FNN 的错误率随 υ 增大而单调上升。当 $\upsilon = 30.09$ 时，FNN 与 TREE 的错误率相同。因此，当 $\upsilon = 30.09$ 时，可得到检验的第一类错误；通过逐渐减小 υ，从 30.09 变为1.0，可绘制检验的功效曲线。

真实世界数据集 2：MNIST 数据集。它是深度学习研究中常用的十分类数据集。keras 库中 MNIST 数据集的标准划分包含一个由 60,000 张图像组成的训练集和一个由 10,000 张图像组成的测试集。我们仅将训练集中的 60,000 张图像用作数据总体。我们对比了两个深度学习模型。

(*A*) 卷积神经网络。所用卷积神经网络的架构表示为 conv 最大池化conv最大池化展平全连接层Softmax。我们使用 ReLU 作为激活函数，使用交叉熵作为目标函数。批量大小和训练轮数分别设为 128 和 250。两个“conv”层中的滤波器数量分别为 32 和 64。两个“conv”层使用相同的卷积核大小 (5，5)。此外，池化大小为 (2，2)

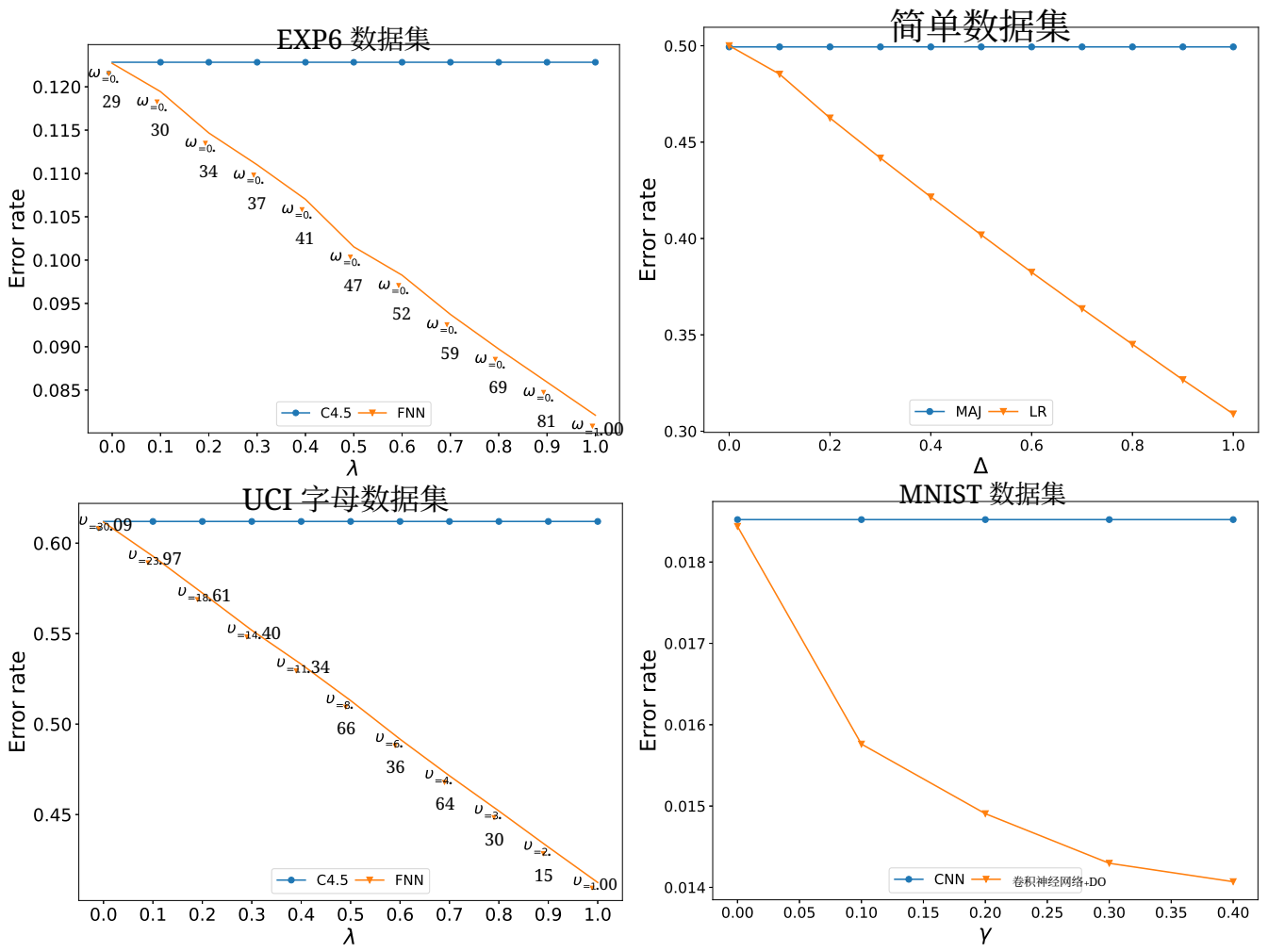

图 3. 算法在合成数据集和真实世界数据集上的真实错误率模拟

使用该池化大小。隐藏全连接层的大小为 500，Softmax 层的大小为 10。

(*B*) 卷积神经网络+DO。 我们对上述卷积神经网络的展平层与全连接层之间的权重应用 Dropout 正则化。使用 Dropout 率 $\gamma \in [0.0, 1.0)$ 来调整卷积神经网络+DO 的错误率。

鉴于训练一个卷积神经网络模型的计算成本很高，我们只生成 100 个数据集，每个数据集的大小为$n = 10,\ 000$，以模拟 100 次独立试验。我们使用整个总体作为测试集，通过对基于这 100 个数据集训练的模型的 100 个估计量取平均值，来计算真实错误率。我们将 Dropout 率从 0.0 调整到 0.4，步长为 0.1。卷积神经网络和 CNN+DO 的真实错误率值绘制在图 3 的最后一幅图中。无论 Dropout 率如何，卷积神经网络的真实错误率均为 1.85%γ。CNN+DO 的真实错误率随着 Dropout 率的增加而降低 γ。当 $\gamma = 0.0$ 时，CNN+DO 的错误率与卷积神经网络相同。随着 γ 增大，CNN+DO 的错误率低于卷积神经网络。因此，对于显著性检验，我们使用 $\gamma = 0.0$ 来获得其第一类错误率，并通过将 γ 从 0.0 调整到 0.4 来获得其功效曲线。

所有显著性检验在模拟数据集和真实世界数据集上的第一类错误列于表 5 中。低于 $\alpha = 0.05$ 的第一类错误以粗体显示。得到两个观察结果。

(1) 5×2 分块交叉验证麦克尼马尔检验和朴素·折 K交叉验证麦克尼马尔检验的所有第一类错误均低于 0.05。这表明这两种检验能够有效减少算法比较中的假阳性结论。分块 3×2 交叉验证 t检验和校准5×2 分块交叉验证 F检验的大多数第一类错误都在合理范围内。相比之下，尽管K折交叉验证配对 t检验和 K折交叉验证-CI Z检验在 Epsilon数据集上的第一类错误合理，但这两种检验在其他数据集上的所有第一类错误均超过 0.05。因此，这两种检验不适合用于比较两种分类算法。此外，RHO 配对 t检验和组合 5×2 交叉验证 t 检验的所有第一类错误均大于 0.05。因此，这两种容易产生假阳性结论的检验在实际应用中应谨慎使用。

（2）在这五个数据集上，我们提出的 5×2 BCV 麦克尼马尔检验几乎取得了最小的第一类错误，且明显低于 0.05。因此，由于公式（10）中 ρ_1 和 ρ_2 的界限较宽松，5×2 BCV 麦克尼马尔检验比其他检验更保守。因此，如何细化 ρ_1 和 ρ_2 的界限以进一步改进麦克尼马尔检验，是未来一项重要的研究方向。

表 5 不同算法比较方法在合成数据集和真实世界数据集上的第一类错误。

族系	Test	合成的			真实世界	
		Epsilon	EXP6	简单	UCI 字母数据集	MNIST
t-检验系列	RHO配对 t检验	0.478	0.272	0.312	0.385	0.250
	K-折交叉验证配对 t-检验	**0.043**	0.076	0.109	0.142	0.090
	5×2 交叉验证配对 t-检验	**0.034**	**0.050**	0.084	0.061	0.080
	校正后的 RHO t-检验	0.053	**0.040**	**0.047**	0.075	**0.000**
	修正后的 10×10 交叉验证 t检验	**0.035**	**0.040**	0.063	0.082	0.080
	组合式 5×2 交叉验证 t检验	0.291	0.156	0.214	0.236	0.150
	分块式 3×2 交叉验证 t检验	0.087	**0.005**	**0.015**	**0.013**	**0.020**
F检验族	组合式 5×2 交叉验证 F检验	**0.028**	**0.028**	0.060	0.057	**0.010**
	校准后的 5×2 分块交叉验证 F检验	**0.035**	**0.025**	0.055	**0.042**	**0.040**
Z-检验族	比例检验	0.056	**0.016**	**0.014**	0.054	**0.020**
	K折交叉验证-置信区间 Z检验	**0.046**	0.062	0.106	0.080	0.090
McNemar 检验族	传统 HO McNemar 检验	**0.031**	**0.037**	**0.029**	0.062	0.060
	朴素 K 折交叉验证麦克尼马尔检验	**0.000**	**0.006**	**0.020**	**0.039**	**0.020**
	5×2 分块交叉验证 麦克尼马尔检验	**0.025**	**0.006**	**0.005**	**0.015**	**0.010**

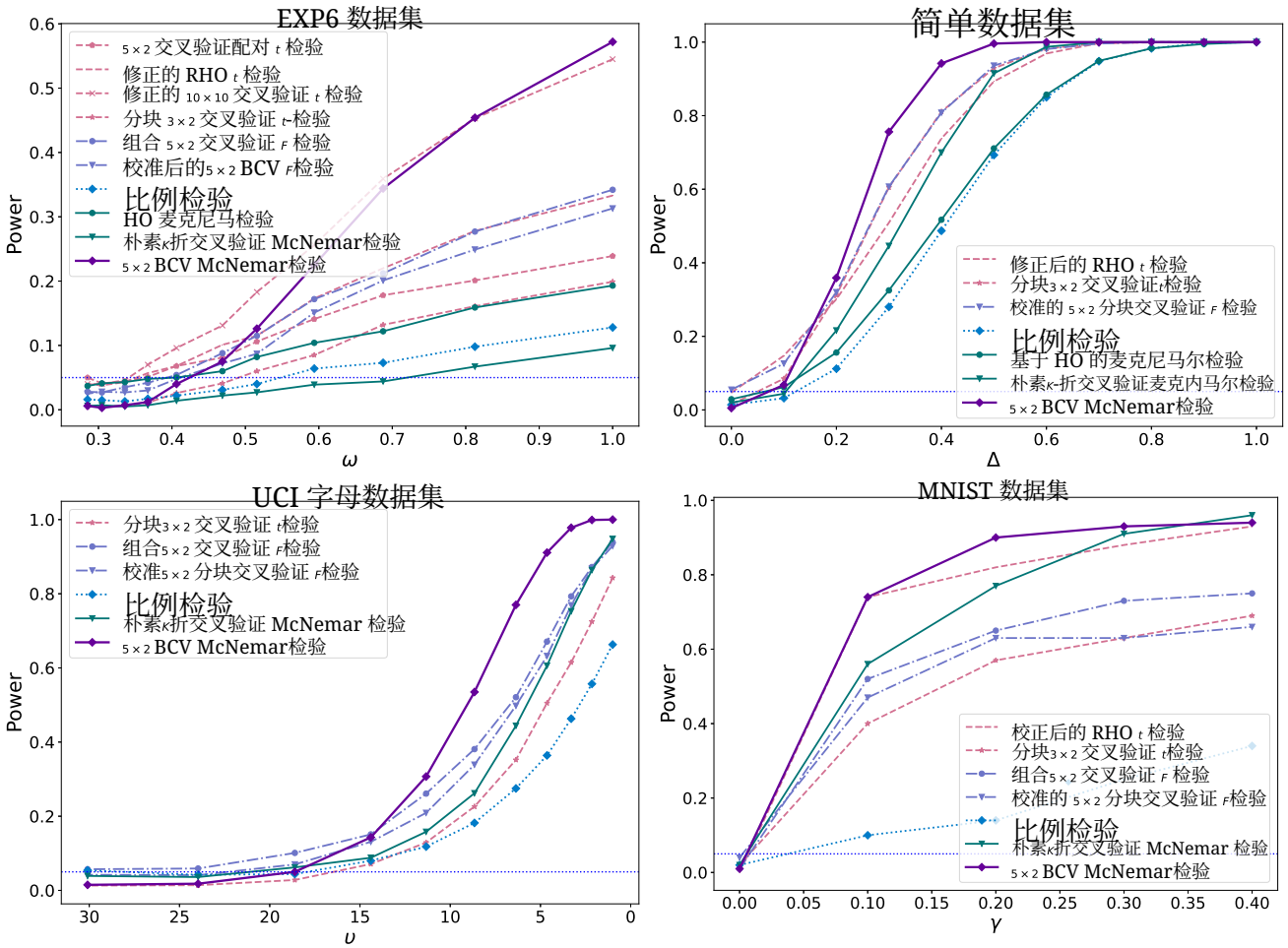

图 4. 不同检验在合成数据集和真实世界数据集上的功效曲线。

图 4 展示了在实验数据集上第一类错误合理的各项检验的功效曲线。从图 4 的所有图中可以得出一个共同结论：与其他检验相比， 5×2 BCV 麦克尼马尔检验的功效曲线更加偏斜。这一观察表明，在比较两个类别时， 5×2 BCV 麦克尼马尔检验在所有检验中检验功效最高。

分类算法。

总之，本节的所有实验结果表明， 5×2 分块交叉验证麦克尼马尔检验不仅第一类错误率相当低，而且检验功效也令人期待。因此， 5×2 分块交叉验证麦克尼马尔检验优于其他检验，并且能够减少算法比较任务中的假阳性结论。因此，我们推荐 5×2

在实际比较两种分类算法的任务中使用分块交叉验证麦克尼马尔检验。

6 结论

在本研究中，我们提出了一种用于比较两种分类算法的 5×2 分块交叉验证麦克尼马尔检验。所提出的检验基于“有效列联表”这一概念构建，该概念从贝叶斯视角有效地压缩了 5×2 分块交叉验证中的十个相关列联表。我们对这些列联表之间的相关性进行了充分研究，并证明了相关性的理论界。大量实验表明，与现有的算法比较方法相比，5×2 分块交叉验证麦克尼马尔检验具有合理的第一类错误和令人期待的检验功效。

未来，我们将改进序贯设置下的$m \times 2$ 分块交叉验证麦克尼马尔检验，并进一步从贝叶斯视角研究 $m \times 2$ 分块交叉验证麦克尼马尔检验，以便在比较分类算法的任务中推断有信息量的决策。

致谢

本研究得到国家自然科学基金（项目编号：61806115）的资助。实验由山西大学高性能计算系统提供支持。

参考文献

[1] J. d. Leeuw、H. Jia、L. Yang、X. Liu、K. S. Schmidt 和 A. K. Skidmore，“比较准确性评估以推断图像分类方法的优劣”，国际遥感杂志，第27卷第1期，223–232页，2006年。[2] T. J. Brinker、A. Hekler、A. H. Enk、C. Berking、S. Hafer- kamp、A. Hauschild、M. Weichenthal、J. Klode、D. Schadendorf、T. Holland-Letz、C. v. Kalle、S. Fr¨ohling、B. Schilling 和 J. S. Utikal，“深度神经网络在黑色素瘤图像分类方面优于皮肤科医生”，欧洲癌症杂志，第119卷，11–17页，2019年。[3] L. Gillick 和 S. J. Cox，“语音识别算法比较中的若干统计问题”，载于国际声学、语音与信号处理会议，1989年，会议论文集，第1卷，532–535页。[4] C. E. Erdem、E. Bozkurt、E. Erzin 和 A. T. Erdem，“基于 RANSAC 的训练数据选择：用于自发语音情感识别”，载于第三届自然环境中的情感交互国际研讨会论文集。ACM出版社，2010年。

[5] [ ] 10 1145 2 1877826 1877831 在线。获取地址：https://doi.org/ 。% F . L. Williams、C. Bannister、M. Arribas-Ayllon、A. Preece 和 I. Spasi´c，“习语在情感分析中的作用”，专家系统与应用，第42卷第21期，7375–7385页，2015年。[6] N. Mukhtar、M. A. Khan 和 N. Chiragh，“在乌尔都语情感分析中有效使用评估指标验证最佳分类器”，认知计算，第9卷第4期，446–456页，2017年。 [在线]。获取地址：https: //doi.org/ 10 1007 12559 017 9481 5 。/s - -

[7] - Y. Sun 、 T. S. Butler 、 A. Shafarenko 、 R. Adams 、 M. Loomes 和 N. Davey，“使用监督分类技术对手写文本进行词语切分”，应用软计算，第7卷第1期，71–88页，2007年。[8] R. Kohavi，“关于交叉验证和自助法在准确率估计与模型选择中的研究”，载于*IJCAI*，第14卷，1995年，会议论文集，1137–1145页。[9] T. G. Dietterich，“比较监督分类学习算法的近似统计检验”，神经计算，第10卷第7期，1895–1923页，1998年。[在线]。获取地址：http://www.mitpressjournals.org/doi/abs/10.1162/089976698300017197

[10]J. D. Rodr´ıguez、A. P´erez 和 J. A. Lozano，“分类误差估计量方差来源统计分析的通用框架”，模式识别，第46卷第3期，855–864页，2013年。 [在线]。获取地址：http://ac.els-cdn.com/S0031320312003998/1-s2.0-S0031320312003998-main.pdf? tid=f45b1bfe-3ad1-11e7-b6c0-00000aacb35e&acdnat=1495006088 – 8c4c1e24a44a58088ac3b782ea57d4f6[11]M. Keller、S. Bengio 和 S. Y. Wong，“非参数统计检验的基准测试”，载于《神经信息处理系统进展》，2006年，会议论文集，651–658页。[12] J. Friedman、T. Hastie 和 R. Tibshirani，《统计学习要素》。Springer统计学丛书，Springer，柏林，2001年，第1卷。[13] J. D. Rodr´ıguez、A. P´erez 和 J. A. Lozano，“预测误差估计中 k 折交叉验证的敏感性分析”，IEEE模式分析与机器智能汇刊，第32卷第3期，569–575页，2010年。[14] R. Wang、Y. Wang、J. Li 和 X. Yang，“泛化误差的块正则化交叉验证估计量”，神经计算，第29卷第2期，519–554页，2017年。[15] E. Alpaydin，“用于比较监督分类学习算法的组合交叉验证 F 检验”，神经计算，第11卷第8期，1885–1892页，1999年。[16] Y. Wang、R. Wang、H. Jia 和 J. Li，“用于比较监督分类学习算法的分块交叉验证 t 检验”，神经计算，第26卷第1期，208–235页，2014年。[在线]。获取地址：http://dblp.uni-trier.de/db/journals/neco/ neco 26 14 .html#WangWJL Y. Wang、J. Li 和 Y. [17] Li，“基于校准平衡交叉验证 F 检验的两种分类学习算法选择”，神经处理快报，第46卷第1期，1–13页，2017年。 [在线]。获取地址：https://doi.org/10.1007/ s11063-016-9569-z[18] C. Nadeau 和 Y. Bengio，“泛化误差的推断”，机器学习，第52卷第3期，239–281页，2003年。[19] Q. McNemar，“关于相关比例或百分比之差的抽样误差的注记”，Psychometrika，第12卷第2期，153–157页，1947年。[20] B. S. Everitt，《列联表分析》。CRC出版社，1992年。[21] C. Goutte 和 E. Gaussier，“精确率、召回率和 F 分数的概率解释及其对评估的启示”，载于《信息检索进展》，D. E. Losada 和 J. M. Fern´andez- Luna 编。柏林，海德堡：Springer Berlin Heidelberg，2005年，345–359页。[22] Y. Wang 和 J. Li，“基于 k 折交叉验证贝塔分布的精确率和召回率可信区间”，神经计算，第28卷第8期，1694–1722页，2016年。[23] R. Wang 和 J. Li，“用于比较两种自然语言处理模型的精确率、召回率和F1度量的贝叶斯检验”，载于计算语言学协会第57届年会论文集，2019年，4135–4145页。[24] O. T. Yildiz，“使用一种新型成对统计检验进行全变量规则归纳”，IEEE知识与数据工程汇刊，第25卷第9期，2105–2118页，2013年。[25] C. J. Wu 和 M. S. Hamada，《实验：规划、分析与优化》。John Wiley & Sons，2011年，第552卷。[26] R. R. Bouckaert 和 E. Frank，《评估用于比较学习算法的显著性检验的可重复性》。Springer，2004年，3–12页。[27] P. Bayle、A. Bayle、L. Janson 和 L. Mackey，“测试误差的交叉验证置信区间”，载于《神经信息处理系统进展》，第33卷，2020年。[28] E. B. Kong 和 T. G. Dietterich，“纠错输出编码可纠正偏差和方差”，载于 ICML’95 第十二届国际机器学习会议论文集，1995年，313–321页。