基于正则化 KL 距离的交叉验证折数 K 的选择

褚荣燕[1],王 钰[2,3*],杨杏丽[1],李济洪[3]
(1. 山西大学 数学科学学院,山西 太原 030006;
2. 山西大学 现代教育技术学院,山西 太原 030006;
3. 山西大学 软件学院,山西 太原 030006)

摘 要:在机器学习中,K 折交叉验证方法常常通过把数据分成多个训练集和测试集来进行模型评估与选择,然而其折数 K 的选择一直是一个公开的问题。注意到上述交叉验证数据划分的一个前提假定是训练集和测试集的分布一致,但是实际数据划分中,往往不是这样。因此,可以通过度量训练集和测试集的分布一致性来进行 K 折交叉验证折数 K 的选择。直观地,KL(Kullback-Leibler)距离是一种合适的度量方法,因为它度量了两个分布之间的差异。然而直接基于 KL 距离进行 K 的选择时,从多个数据实验结果发现随着 K 的增加 KL 距离也在增大,显然这是不合适的。为此,提出了一种基于正则化 KL 距离的 K 折交叉验证折数 K 的选择准则,通过最小化此正则 KL 距离来选择合适的折数 K。进一步多个真实数据实验验证了提出准则的有效性和合理性。

关键词:K 折交叉验证;折数 K 的选择;KL(Kullback-Leibler)距离;正则化;机器学习
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2021)03-0052-06
doi:10.3969/j.issn.1673-629X.2021.03.009

A Selection Criterion of Fold K in Cross-validation Based on Regularized KL Distance

CHU Rong-yan[1], WANG Yu[2,3*], YANG Xing-li[1], LI Ji-hong[3]
(1. School of Mathematical Sciences, Shanxi University, Taiyuan 030006, China;
2. School of Modern Educational Technology, Shanxi University, Taiyuan 030006, China;
3. School of Software, Shanxi University, Taiyuan 030006, China)

Abstract: In machine learning, the K-fold cross-validation method often divides the data into multiple training and test sets for model evaluation and selection. However, the selection of the fold K is always an open problem. Note that one of the premises of the above cross-validation data division assumes that the training set and the test set have the same distribution, but in actual data division, this is often not the case. Therefore, the selection of the fold K can be performed by measuring the distribution consistency of the training set and the test set in K-fold cross-validation. Intuitively, KL (Kullback-Leibler) distance is a suitable measure because it measures the difference between two distributions. However, when selecting K directly based on the KL distance, it is found from multiple data experimental results that the KL distance also increases with the increase of K, which is obviously inappropriate. To this end, a selection criterion of the fold K in K-fold cross-validation based on regularized KL distance is proposed, and the appropriate fold K is selected by minimizing this regular KL distance. Multiple real data experiments in a recent step have verified the effectiveness and rationality of the proposed criterion.

Key words: K-fold cross-validation; selection of the fold K; KL distance (Kullback-Leibler distance); regularized; machine learning

0 引 言

在机器学习中,交叉验证技术广泛地应用于模型(算法)性能评估、特征选择、模型选择、模型参数确定、过拟合检验等任务[1-5]。例如,给定一组线性空间(模型),在这些线性空间(模型)中选择最佳的最小二乘估计量时可用交叉验证进行模型选择。所谓交叉验证技术,即数据集被随机地切分为多个训练集和测试集,训练集用来进行模型的拟合,测试集用来进行模型

收稿日期:2020-05-16 **修回日期**:2020-09-18
基金项目:山西省应用基础项目研究计划(201901D111034,201801D211002);国家自然科学基金资助项目(61806115)
作者简介:褚荣燕(1995-),女,硕士研究生,研究方向为统计机器学习;通信作者:王 钰(1981-),男,副教授,硕导,博士,研究方向为统计机器学习、数据挖掘、图像处理等。

性能的评估,最后通过多次性能评估的均值(投票)来分析模型的优劣。其中,常见的交叉验证技术包括留一(leave-one-out)交叉验证、Hold-out 交叉验证、RLT(repeated learning testing)交叉验证、5×2 交叉验证、组块 3×2 交叉验证、组块 $m \times 2$ 交叉验证、K 折交叉验证等[6-9]。在这些交叉验证技术中,K 折交叉验证是最广泛使用的方法,因为它依赖于一个整数参数 K,比其他经典交叉验证方法的计算代价更小。K 折交叉验证指的是数据集被平均分成 K 个大小近似相同但不相交的子集,选取其中 $K-1$ 个子集作为训练集来拟合模型,剩下的一个子集作为测试集来评估模型性能。然而,在机器学习中,关于 K 折交叉验证方法的折数 K 的选择虽然很多文献中都对其进行了研究并给出了推荐,但它一直是一个公开未解决的问题[10-15]。比如文献[6]中提到当模型选择的目标是估计时,最优的折数 K 为 5 到 10 之间,这是因为 K 值越大,统计性能不会增加太多,并且小于 10 次分割的平均值在计算上仍然可行。文献[8,11,12,14]皆推荐在进行泛化误差估计,算法性能对照和超参数选择时应选择二折或多次二折重复的交叉验证(5×2 交叉验证、组块 3×2 交叉验证、组块 $m \times 2$ 交叉验证)。文献[15]在多个分类器下的大量实验中验证了在进行模型精度估计和模型选择时十折交叉验证优于留一交叉验证。另外,注意到上述文献中大多是从模型性能评估和选择的角度来进行折数的选择,但事实上,执行上述交叉验证过程的一个前提条件是训练样本和测试样本的分布一致,然而在实际中对于训练样本和测试样本的分布是否一致许多文献中并没有验证。因此,该文考虑通过度量训练样本和测试样本的分布一致性来进行 K 折交叉验证折数 K 的选择。

实际中,常用的度量两个分布函数之间差异的度量有 KL(Kullback-Leibler)距离、全变差(total variation)距离、海灵格(Hellinger)距离、KMM(kernel mean matching)度量、MMD(maximum mean discrepancy)度量、Wasserstein 距离等[16-17],其中 KL 距离是最简单且广泛使用的方法。因此,该文基于 KL 距离进行 K 折交叉验证中折数 K 的选择。

为此在 UCI 数据库中选取的四个数据集上进行了实验,实验发现直接用 KL 距离来进行 K 折交叉验证中折数 K 的选择可能是不合理的,因为如图 1 所示,随着 K 折交叉验证折数 K 的增加,训练样本和测试样本分布间 KL 距离也在增大,这样直接基于 KL 距离进行折数 K 的选择往往选出的都是最小的或接近最小的折数 K。因此不能直接应用 KL 距离来进行 K 折交叉验证折数 K 的选择,为此考虑通过对 KL 距离增加一个随着折数 K 增加而变小的正则化项来得到一个正则化的 KL 距离,以此作为交叉验证中折数 K 的选择准则。

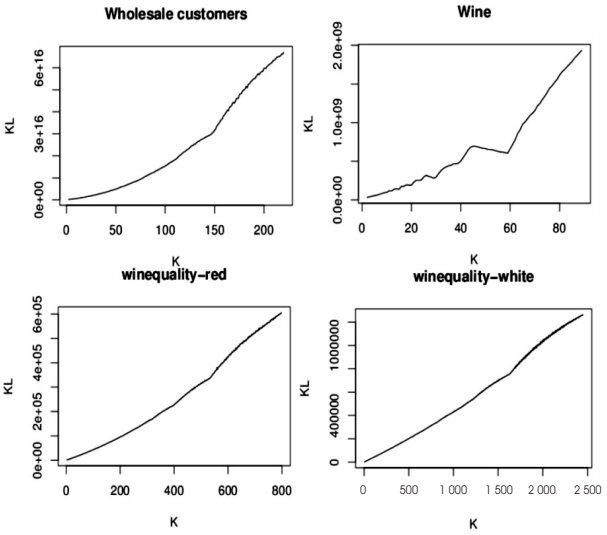


图 1　随着折数 K 的变化 KL 距离的变化

1　KL 距离

在两个分布函数之间差异的度量中最广泛使用的方法是 KL 距离。接下来给出 KL 距离的定义。

如果记 $P(x)$ 和 $Q(x)$ 是两个已知的分布函数,则 $P(x)$ 和 $Q(x)$ 之间的 KL 距离为:

$$D_{\mathrm{KL}}[P(x)\,||\,Q(x)] = \int_D P(x)\log\frac{P(x)}{Q(x)}\mathrm{d}x \tag{1}$$

特别地,当 $P(x)$ 和 $Q(x)$ 分别为高斯分布 $N_d(\boldsymbol{u}_{\mathrm{S}}, \boldsymbol{\Sigma}_{\mathrm{S}})$ 和 $N_d(\boldsymbol{u}_{\mathrm{F}}, \boldsymbol{\Sigma}_{\mathrm{F}})$ 时,根据矩阵的性质及多元高斯分布期望和协方差的性质[18],其 KL 距离可写为:

$$\begin{aligned} D_{\mathrm{KL}}[P(x)\,||\,Q(x)] &= \int_D P(x)\log\frac{P(x)}{Q(x)}\mathrm{d}x = \\ &\frac{1}{2}(n_2-n_1)\log 2\pi + \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_{\mathrm{F}}|}{|\boldsymbol{\Sigma}_{\mathrm{S}}|} + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{S}})\right] + \\ &\frac{1}{2}\left[(\boldsymbol{u}_{\mathrm{S}}-\boldsymbol{u}_{\mathrm{F}})'\boldsymbol{\Sigma}_{\mathrm{T}}^{-1}(\boldsymbol{u}_{\mathrm{S}}-\boldsymbol{u}_{\mathrm{F}})\right] - \frac{1}{2}d = \\ &\frac{1}{2}(n_2-n_1)\log 2\pi + \frac{1}{2}B_{\boldsymbol{\Sigma}_F^{-1}} + \frac{1}{2}M_{\boldsymbol{\Sigma}_F^{-1}} - \frac{1}{2}d \end{aligned} \tag{2}$$

其中,$B_{\boldsymbol{\Sigma}_F^{-1}} = \log\frac{|\boldsymbol{\Sigma}_{\mathrm{F}}|}{|\boldsymbol{\Sigma}_{\mathrm{S}}|} + \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{S}})$,$M_{\boldsymbol{\Sigma}_F^{-1}} = (\boldsymbol{u}_{\mathrm{S}}-\boldsymbol{u}_{\mathrm{F}})'\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}(\boldsymbol{u}_{\mathrm{S}}-\boldsymbol{u}_{\mathrm{F}})$ 是马氏距离,d 是数据的维数。

具体地,在对数据进行分析时,$P(x)$ 和 $Q(x)$ 的总体均值和总体协方差一般使用样本均值和样本协方差来估计[19]:

$$\begin{aligned} \hat{\boldsymbol{u}}_{\mathrm{S}} &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_i \\ \hat{\boldsymbol{\Sigma}}_{\mathrm{S}} &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\boldsymbol{u}_{\mathrm{S}})(x_i-\boldsymbol{u}_{\mathrm{S}})' \end{aligned} \tag{3}$$

$$\hat{\boldsymbol{u}}_{\mathrm{F}}=\frac{1}{n_2}\sum_{i=1}^{n_2}x_i^*$$

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{F}}=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_i^*-\boldsymbol{u}_{\mathrm{F}})(x_i^*-\boldsymbol{u}_{\mathrm{F}})' \tag{4}$$

此时,把式(3)和式(4)代入式(2)中,得到式(5):

$$\begin{aligned}D_{\mathrm{KL}}[P(x)\,||\,Q(x)]=&\frac{1}{2}(n_2-n_1)\log 2\pi+\\&\frac{1}{2}\Big[\log\frac{|\hat{\boldsymbol{\Sigma}}_{\mathrm{F}}|}{|\hat{\boldsymbol{\Sigma}}_{\mathrm{S}}|}+\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\mathrm{F}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathrm{S}})\Big]+\\&\frac{1}{2}\Big[(\hat{\boldsymbol{u}}_{\mathrm{S}}-\hat{\boldsymbol{u}}_{\mathrm{F}})'\hat{\boldsymbol{\Sigma}}_{\mathrm{F}}^{-1}(\hat{\boldsymbol{u}}_{\mathrm{S}}-\hat{\boldsymbol{u}}_{\mathrm{F}})\Big]-\frac{1}{2}d=\\&\frac{1}{2}(n_2-n_1)\log 2\pi+\frac{1}{2}B_{\hat{\boldsymbol{\Sigma}}_{\mathrm{F}}^{-1}}+\\&\frac{1}{2}M_{\hat{\boldsymbol{\Sigma}}_{\mathrm{F}}^{-1}}-\frac{1}{2}d\end{aligned} \tag{5}$$

2 K 折交叉验证折数 K 的选择准则

本节我们将基于上一节定义的 KL 距离进行具体分析,通过对原始 KL 距离增加一正则化项,利用此正则化 KL 距离来选择使得训练样本和测试样本分布尽可能一致的 K 折交叉验证的折数 K。

2.1 K 折交叉验证折数 K 的选择准则

具体地,如果定义训练集为 $D_S=\{(x_i,y_i)\}_{i=1}^{n_1}$,其中 $x_i\in\Re^d$ 是 d 维输入空间,y_i 是输出空间。测试集为 $D_{\mathrm{F}}=\{(x_i^*,y_i^*)\}_{i=1}^{n_2}$,其中 $x_i^*\in\Re^d$ 是 d 维输入空间,y_i^* 是输出空间。训练样本和测试样本的分布分别为 $P(x)$ 和 $Q(x)$。由上一节中 KL 距离的定义知,训练样本和测试样本之间的 KL 距离即为式(5)。

观察式(5)发现 $D_{\mathrm{KL}}[P(x)\,||\,Q(x)]\neq D_{\mathrm{KL}}[Q(x)\,||\,P(x)]$,也就是说 KL 距离不是对称的,因此该文考虑对称的 KL 距离,即:

$$\begin{aligned}D_{\mathrm{SKL}}[P(x),\,Q(x)]=&\frac{1}{2}[D_{\mathrm{KL}}[P(x)\,\|\,Q(x)]+\\&D_{\mathrm{KL}}[Q(x)\,\|\,P(x)]]=\frac{1}{4}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{S}}+\\&\boldsymbol{\Sigma}_{\mathrm{S}}^{-1}\boldsymbol{\Sigma}_{\mathrm{F}})+\frac{1}{4}M_{\boldsymbol{\Sigma}_{\mathrm{S}}^{-1}+\boldsymbol{\Sigma}_{\mathrm{F}}^{-1}}(\boldsymbol{u}_{\mathrm{S}},\boldsymbol{u}_{\mathrm{F}})-\frac{1}{2}d\end{aligned} \tag{6}$$

进一步,基于 K 折交叉验证 K 次重复的对称 KL 距离为:

$$D_{\mathrm{AKL}}[P(x),Q(x)]=\frac{1}{K}\sum_{k=1}^{K}D_{\mathrm{SKL}}^{(k)}[P(x),Q(x)] \tag{7}$$

其中,$D_{\mathrm{SKL}}^{(k)}[\cdot]$ 定义为第 K 折训练样本与测试样本分布之间的 KL 距离,K 为折数。

虽然,在经验上,基于训练样本与测试样本的 KL 距离选择合适的折数 K 的方法是一个比较理想的方法,但是在实际结果中发现直接基于式(7)进行 K 折交叉验证折数 K 的选择并不是一个好的方法(详见图 1),因为随着折数 K 的增加 KL 距离也会增大,这会导致几乎所有数据选出的折数都较小。为了解决此问题,该文进一步提出了一种新的基于正则化 KL 距离的 K 折交叉验证折数 K 的选择准则。

2.2 正则化 KL 距离的 K 折交叉验证折数 K 的选择

把一个随着折数 K 增加而减小的函数作为一个正则化项添加到式(7)中,这样对原始 KL 距离起到折中作用。为此,给出如下正则化 KL 距离选择准则:

$$D_{\mathrm{ReKL}}(K)=D_{\mathrm{AKL}}[P(x),Q(x)]+\lambda f(K) \tag{8}$$

其中,$D_{\mathrm{AKL}}[\cdot]$ 定义为第 K 折训练样本与测试样本之间的平均 KL 距离,K 为折数;$f(K)$ 是关于 K 的函数;λ 是调节参数,通过调整 λ 确定正则化的程度。这时,基于最小化正则化 KL 距离的交叉验证折数 K 的选择准则为:

$$\hat{K}=\arg\min_{K}D_{\mathrm{ReKL}}(K) \tag{9}$$

3 真实数据实验

选取了 UCI 数据库中的 Wholesale customers, Wine, wine quality-red, wine quality-white 四个数据集来进行实验,以验证提出的基于正则化 KL 距离的 K 折交叉验证折数 K 的选择准则的合理性和有效性。

3.1 实验设置

四个数据集的相关描述如下:

(1)Wholesale customers 数据集:关于批发商批发产品年度支出的一个二类分类(三类分类)数据集,根据批发商渠道把其中的餐饮业渠道分为第一类,零售渠道分为第二类(或按照客户所在的区域(里斯本,波尔图,其他区域这三个区域)把数据分为三类),包含 7 个特征(Fresh, Milk, Grocery, Frozen 等特征),共 440 个样本。

(2)Wine 数据集:关于三种不同品种的葡萄酒化学分析的一个三类分类数据集,三种葡萄酒即为三类,含有 13 个特征(Malic acid, Ash, Alkalinity of ash, Magnesium 等特征)且每个特征是连续的变量,共 178 个样本。

(3)wine quality-red 数据集:关于红葡萄酒质量优劣检测的一个二类分类(六类分类)数据集,根据酒的质量把质量指标大于 5 的分成一类,质量指标小于等于 5 的分成另一类(或不同的指标各表示一类),包含 11 个特征(fixed acidity, volatile acidity, citric acid, residual sugar 等特征),共 1 599 个样本。

(4)wine quality-white 数据集:关于白葡萄酒质量优劣检测的一个二类分类(七类分类)数据集,根据酒的质量把质量指标大于 5 的分成一类,质量指标小于等于 5 的分成另一类(或不同的指标各表示一类),包含 11 个特征(fixed acidity, volatile acidity, citric acid,residual sugar 等特征),共 4 898 个样本。

此外为了考虑不同类别对于 KL 距离的影响,按照数据的属性描述重新进行了分类。例如 Wholesale customers 数据根据渠道和区域分为 3 类和 2 类,wine quality-red 数据根据质量指标分为 6 类和 2 类,wine quality-white 数据根据质量指标分为 7 类和 2 类。

在式(8)给出的基于正则化 KL 距离的选择准则中,正则化函数设置为 $f(K)=\exp(-K)$。这是因为在如 Wholesale customers 数据集中,KL 距离是从 2.12×10^{14} 变化到 6.71×10^{16},而折数 K 是从 2 变化到 220,它们是不同数量级的,因此,为了二者的折中,不失一般性,正则化函数设置为 $f(K)=\exp(-K)$。事实上,不同的正则化函数(例如 $\log(x)$ 函数, $\exp(x)$ 函数)对于最优折数 K 的选择是没有影响的。因为对于不同的正则化函数,通过调节不同的调节参数值始终可以选出相同的最优折数 K,只是收敛速度不同罢了[20]。

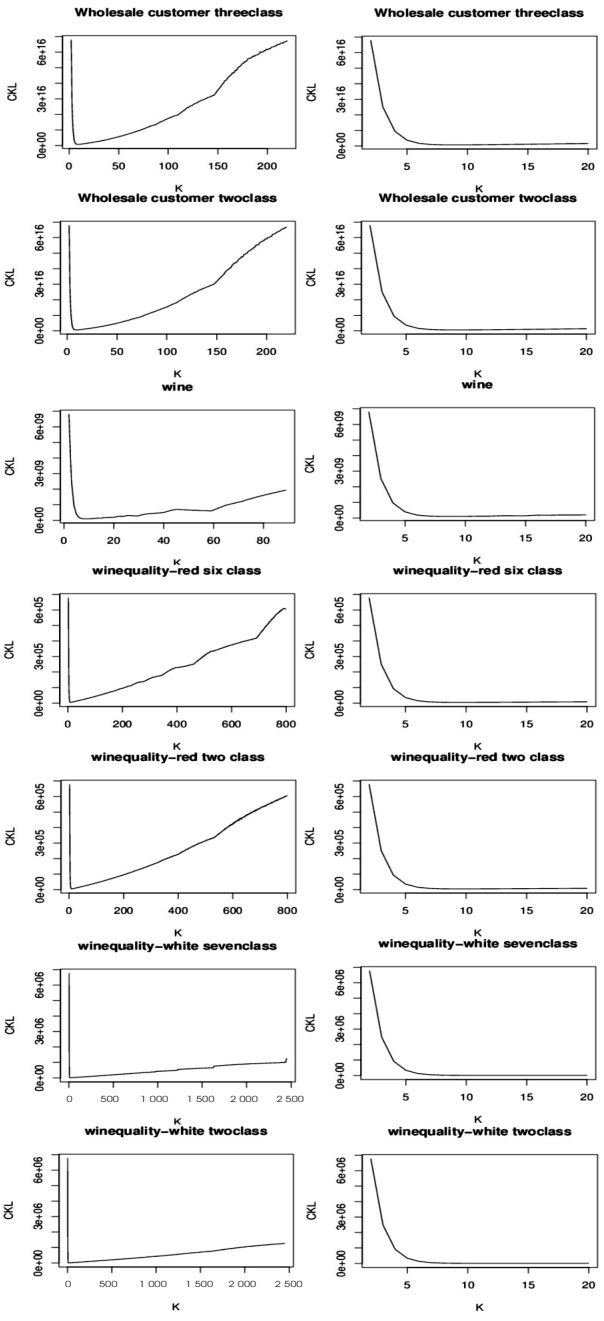
进一步,为了保证实验的准确性,对该实验过程重复 1 000 次,最后取这 1 000 次的平均值作为最终正则化 KL 距离。

注:交叉验证折数 K 的实际取值范围为[2, $n/2$], n 为数据容量,因为在使用多元高斯分布来估计 KL 距离时,训练集/测试集最小需要两个样本估计样本协方差阵。

3.2　实验结果

基于 3.1 节的数据集,首先验证了随着 K 折交叉验证的折数 K 的增加,所有训练样本和测试样本分布之间的 KL 距离逐渐增大。然后,给出了基于提出的正则化 KL 距离的 K 折交叉验证折数 K 的选择准则的选择结果。

图 1 展示了不同折交叉验证的训练样本和测试样本之间的 KL 距离。根据图 1,首先可以看到所有训练样本和测试样本分布之间的差异是明显的,例如,在 Wholesale customers 数据中,折数从 2 折增到 10 折时,对应的 KL 距离从 2.12×10^{14} 上升到 7.04×10^{14},在 Wine 数据中,折数从 2 折增到 10 折时,对应的 KL 距离从 3.31×10^{7} 上升到 9.94×10^{7}。第二,图中直观地显示了随着折数 K 的增加,KL 距离逐渐增大。例如,wine quality-red 数据集中,折数范围从 2 变化到 799 时,KL 距离从 1 532 上升到 6.05×10^{5}。wine quality-white 数据集中,折数范围从 2 变化到 2 449 时,KL 距离从 1 695 上升到 1.26×10^{6},也就是说当折数从 2 上升到 $n/2$(n 为样本量)时,不同数据的 KL 距离都是持续上升的。且几乎所有数据训练样本和测试样本的 KL 距离都是 2 折或者接近 2 折时最小,显然这是不合适的,因为它总选择最小或接近最小的折数,因此为 KL 距离增加一个正则化项,提出了基于正则化 KL 距离的 K 折交叉验证折数 K 的选择准则,通过最小化式(8)中给出的准则来选择最优折数。结果如图 2 所示。



(a)　　(b)

图 2　随着折数 K 的增加正则化 KL 距离的变化
(图(b)是图(a)左下角 2 到 20 折的部分图像)

图 2 展示了正则化后的 $D_{\mathrm{ReKL}}(\cdot)$ 变化趋势。为了更清楚地看到正则化 KL 距离的变化,表 1 ~ 表 4 给出

了 $K=2,5,8,9,10,11,12,13,14,15,20$ 时，四个数据集上正则化 KL 距离的数值。首先可以看出，随着折数的增加正则化 KL 距离先减小后增大，例如在表 1 的 Wholesale customers 数据集中 3 类的情况下，9 折时正则化 KL 距离的值为 7.27×10^{14}，10 折时正则化 KL 距离的值为 7.18×10^{14}，20 折时正则化 KL 距离的值为 15.92×10^{14}。在表 2 的 Wine 数据集中，9 折时正则化 KL 距离的值为 10×10^{7}，10 折时正则化 KL 距离的值为 9.94×10^{7}，20 折时正则化 KL 距离的值为 19.47×10^{7}。在表 4 的 wine quality-white 数据集中，9 折时正则化 KL 距离的值为 9.93×10^{3}，10 折时正则化 KL 距离的值为 6.23×10^{3}，20 折时正则化 KL 距离的值为 7.78×10^{3}，这都验证了提出的方法是合适的。

表 1 Wholesale customers 数据集的 D_{CKL} 距离结果

数据集	折数	3 类 D_{CKL} 距离（$\times10^{14}$）	2 类 D_{CKL} 距离（$\times10^{14}$）
Wholesale customers	2	678.88	678.55
	5	37.24	36.69
	8	7.35	6.43
	9	7.27	***5.99***
	10	**7.18**	6.14
	11	7.90	6.70
	12	8.64	7.14
	13	9.46	7.90
	14	10.42	8.76
	15	11.11	9.35
	20	15.92	13.26

表 2 Wine 数据集的 D_{CKL} 距离结果

数据集	折数	3 类 D_{CKL} 距离（$\times10^{7}$）
Wine	2	680.28
	5	39.31
	8	9.99
	9	10.00
	10	**9.94**
	11	11.43
	12	11.52
	13	14.08
	14	14.27
	15	14.26
	20	19.47

另外，可以在表 1 ~ 表 4 中看到每个数据的最优折数（表 1 ~ 表 4 中的黑色粗斜体表示不同数据分为 2 类时最优的折数及其对应的正则化 KL 距离值，黑色粗体表示不同数据分为其他类时最优的折数及其对应的正则化 KL 距离值），例如 Wholesale-customers 数据分 3 类时最优折数为 10 折，分 2 类时最优折数为 9 折。Wine 数据最优折数为 10 折。wine quality-red 数据分 6 类时最优折数 10 折，2 类时最优折数为 9 折。wine quality-white 数据分为 7 类时最优折数为 12 折，2 类时最优折数为 12 折。最后，发现不同分类类别对同一数据集的训练样本和测试样本之间的正则化 KL 距离是有影响的，一般情况下同一数据集下多类分类比二类分类的正则化 KL 距离的差异小（除 Wholesale customers 数据集外），例如在表 4 中，当折数为 12 时，7 类的正则化 KL 距离的数值约为 5.17×10^{3}，2 类的正则化 KL 距离的数值约为 5.25×10^{3}。

表 3 wine quality-red 数据集的 D_{CKL} 距离结果

数据集	折数	6 类 D_{CKL} 距离（$\times10^{3}$）	2 类 D_{CKL} 距离（$\times10^{3}$）
wine quality-red	2	678.09	678.27
	5	36.02	36.06
	8	5.06	5.24
	9	4.41	***4.52***
	10	**4.40**	4.55
	11	4.75	4.81
	12	4.95	5.06
	13	5.32	5.50
	14	5.74	5.86
	15	6.13	6.25
	20	8.08	8.10

表 4 wine quality-white 数据集的 D_{CKL} 距离结果

数据集	折数	7 类 D_{CKL} 距离（$\times10^{3}$）	2 类 D_{CKL} 距离（$\times10^{3}$）
wine quality-white	2	6 768.23	6 768.27
	5	339.12	339.25
	8	20.12	20.18
	9	9.93	10.01
	10	6.23	6.41
	11	5.39	5.52
	12	**5.17**	***5.25***
	13	5.32	5.29
	14	5.55	5.74
	15	5.81	6.02
	20	7.78	8.01

4 结束语

K 折交叉验证技术在机器学习和统计学中被广泛使用，但关于其折数 K 的选择一直是一个公开未解决

的问题。而在传统的机器学习中使用 K 折交叉验证进行分析时,都是在假设训练样本和测试样本分布一致进行的,但是实际中训练样本和测试样本的分布往往不一致。该文在考虑 K 折交叉验证过程中训练样本和测试样本的分布一致性的情况下,利用 KL 距离来度量训练集样本和测试集样本二者分布的差异,通过实验发现直接使用 KL 距离进行选择时,往往选出的是最小的或者接近最小的折数,这并不合理。基于此,提出了一种基于正则化 KL 距离的 K 折交叉验证折数 K 的选择准则。并通过 UCI 数据库中的四个数据集对提出的准则进行验证,最终结果验证了该选择准则的合理性和有效性。

同时我们应该看到,虽然我们通过使训练样本和测试样本的分布差异最小化选出了 K 折交叉验证的合适的折数 K,但是训练样本和测试样本的分布之间仍存在差异,本文并未给出使训练集与测试集尽可能一致的校正策略,未来我们将在此框架下进一步研究二者分布的校正策略。

参考文献:

[1] GRANDVALET Y,BENGIO Y. Hypothesis testing for cross validation[D]. Montreal:University of Montreal,2006.

[2] RODRÍGUEZ J D,PEREZ A P,LOZANO J A. Sensitivity analysis of k-fold cross validation in prediction error estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(3):569-575.

[3] 王瑞波,王　钰,李济洪. 面向文本数据的正则化交叉验证方法[J]. 中文信息学报,2019,33(5):54-65.

[4] LEVER J,KRZYWINSKI M,ALTMAN N. Points of significance:model selection and overfitting[J]. Nature Methods,2016,13(9):703-704.

[5] 杨　柳,王　钰. 泛化误差的各种交叉验证估计方法综述[J]. 计算机应用研究,2015,32(5):1287-1290.

[6] ARLOT S,CELISSE A. A survey of cross-validation procedures for model selection[J]. Statistics Surveys,2010,4:40-79.

[7] WANG Yu,LI Jihong,LI Yanpeng. Measure for data partitioning in m×2 cross-validated[J]. Pattern Recognition Letters,2015,65:211-217.

[8] ALPAYDIN E. Combined 5×2cv F-test for comparing supervised classification learning algorithms[J]. Neural Computation,1999,11(8):1885-1892.

[9] FUSHIKI T. Estimation of prediction error by using K-fold cross-validation[J]. Statistical Computer,2011,21:137-146.

[10] 杨　稳,刘晓宁,朱　菲. 基于支持向量机的颅骨性别识别[J]. 计算机技术与发展,2019,29(2):43-47.

[11] WANG Yu,WANG Ruibo,JIA Huichen,et al. Blocked 3×2 cross-validated t-test for comparing supervised classification learning algorithms[J]. Neural Computation,2014,26(1):208-235.

[12] WANG Ruibo,WANG Yu,LI Jihong,et al. Block-regularized m×2 cross-validated estimator of the generalization error[J]. Neural Computation,2017,29:519-554.

[13] 同　鸣,王　凡,王　硕,等. 一种 3D HOGTCC 和 3D HOOFG 的行为识别新框架[J]. 计算机研究与发展,2015,52(12):2802-2812.

[14] BURMAN P. A comparative study of ordinary cross-validation,v-fold cross-validation and the repeated learning-testing methods[J]. Biometrika,1989,76(3):503-514.

[15] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//International joint conference on artificial intelligence. San Francisco:Morgan Kaufmann Publishers Inc,1995:1137-1143.

[16] IYER A,NATH S,SARAWAGI S. Maximum mean discrepancy for class ratio estimation:convergence bounds and kernel selection[C]//International conference on machine learning (ICML). Beijing,China:ACM,2014:I-530-I-538.

[17] SZYMON M W. Metric diff union along foliations[M]. Switzerland:Springer Nature,2017:1-5.

[18] 茆诗松,王静龙,濮晓龙. 高等数理统计[M]. 第 2 版. 北京:高等教育出版社,2006:15-20.

[19] CHEN B,LAM W,TSANG I,et al. Location and scatter matching for dataset shift in text mining[C]//IEEE international conference on data mining. Sydney,NSW,Australia:IEEE,2011:773-778.

[20] WANG Yu,WANG Chunheng,SHI Cunzhao,et al. A selection criterion for the optimal resolution of ground-based remote sensing cloud images for cloud classification[J]. IEEE Transactions on Geoscience and Remote Sensing,2019,57(3):1358-1367.