第49卷第12期 中国科学技术大学学报 Vol. 49, No. 12
2019年12月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Dec. 2019

文章编号:0253-2778(2019)12-0974-11

高维数据情形下的一种基于随机投影的集成分类方法

崔文泉,黄禹侨

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:针对高维数据的分类问题,提出一种基于随机投影的决策树集成学习方法(Projection Forest,简记 PJForest).该方法以决策树为基分类器,利用一系列随机投影对数据进行降维,基于降维后的数据构建相应的一系列决策树,而后通过集成学习构造集成分类器.利用适当的随机投影对数据进行降维,能保持数据几何结构的信息;且通过随机投影对原始数据进行扰动,能丰富决策树的多样性,经过适当集成可有效克服噪音的影响,进而提升 PJForest 的泛化能力.证明了 PJForest 泛化误差的极限性质,得到泛化误差在一定意义下的收敛速度.还开展大量的模拟研究,并对实际数据进行了实证分析.模拟研究的结果表明,PJForest 能有效地对包含大量噪音的高维数据进行分类,与已有的诸如随机森林、Xgboost 这些方法相比,有更好的分类性能.

关键词:决策树;多样性;高维;分类;集成学习;随机投影

中图分类号:O212.7 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.12.004

2010 Mathematics Subject Classification:Primary 62G99; Secondary 62-07

引用格式:崔文泉,黄禹侨. 高维数据情形下的一种基于随机投影的集成分类方法[J]. 中国科学技术大学学报,2019,49(12):974-984.

CUI Wenquan, HUANG Yuqiao. A new random projection-based ensemble classifier for high-dimensional data[J]. Journal of University of Science and Technology of China, 2019,49(12):974-984.

A new random projection-based ensemble classifier for high-dimensional data

CUI Wenquan, HUANG Yuqiao

(*Department of Statistics and Finance, Shool of Management, University of Science and of Technology of China, Hefei* 230026, *China*)

Abstract: A decision tree ensemble method based on random projection(projection forest, PJForest) was proposed to solve the classification problem of high-dimensional data. This method used the decision tree as the base classifier and reduced the dimensionality of the data by using a series of random projections. Then based on dimensionally reduced data, a series of decision trees were constructed, and then the ensemble classifier was constructed through ensemble learning. Using appropriate random projection to reduce the dimensionality of the data can preserve the information contained in the geometric structure of the data. Moreover, perturbation of raw data through random projection can enrich the diversity of decision trees. After proper ensemble learning, it can effectively overcome the influence of noise and improve the generalization ability of PJForest. The limiting property of PJForest generalization error was proved and the convergence rate of generalization error under certain conditions was obtained. Many simulation

收稿日期:2019-04-14;**修回日期**:2019-05-23

基金项目:国家自然科学基金(71873128),安徽省自然科学基金(1308085MA02)资助.

作者简介:崔文泉(通讯作者),男,1964年生,博士/副教授. 研究方向:数理统计. E-mail:wqcui@ustc.edu.cn

studies were conducted and empirical studies on real life data were empirically analyzed. The simulation results showed that the method of PJForest can effectively classify high dimensional data with a large amount of noises, and has better properties than current classification methods such as random forest, Xgboost.

Key words: decision tree; diversity; high-dimensional classification; ensemble learning; random projection

0　引言

随着时代的发展与科技的进步,信息呈爆炸式增长,高维数据越来越频繁地出现在各个领域,比如信号数据、图像数据、金融数据等,有许多实际中的应用要求人们对高维数据进行分类.研究如何有效地对高维数据进行分类是一个具有挑战性的问题,处理高维数据时会遇到各种各样的困难,比如说样本量不充足、数据存在大量噪音等[1-4].

决策树是一种常见的机器学习方法,具体实现的算法有许多,分类回归树(classification and regression tree,CART)[5]是其中一种十分重要的方法,在实际中有着广泛的应用,如 Save 等[6]利用决策树进行信用卡欺诈风险的检测;Gokgoz 和 Subasi[7]使用决策树对生物信号进行分类;Salmam 等[8]使用决策树进行人脸识别.

但是,决策树有着易过拟合、不稳定等缺点.为此,许多学者做了大量的改进研究,其中一个重要的方法是集成学习.集成学习是一种结合多个分类器的学习策略,该策略能有效提升单一分类器的分类性能.Breiman[9]首次提出 Bagging 集成方法,在多个 Boostrap 样本集上训练决策树,集成所有决策树并通过投票法进行预测.Ho[10]提出一种基于随机子空间的决策树集成方法,在训练每棵决策树时,随机选择部分特征进行最优特征搜索与分割;Breiman[11]在此基础上进一步提出随机森林方法,在每个 Boostrap 样本上使用随机子空间法构建决策树,最后集成并通过投票法作出预测.他们的研究表明,通过训练多棵决策树进行集成,并在训练基学习器过程中引入某种随机性扰动,如 Boostrap 抽样、随机子空间,能有效克服单个决策树过拟合与不稳定的缺点.有许多实际中的应用表明随机森林能有效处理高维分类问题,如 Diaz-Uriarte 和 De Andres[12]使用随机森林对基因数据进行分类,Joelsson 等[13]使用随机森林对高光谱数据进行分类.

但是,随机森林方法也有局限性,其在处理包含大量噪音的高维数据时分类性能较差[14].这是因为随机森林使用随机子空间法构建决策树,对特征的随机抽样常常会选择到噪音特征,噪音会进一步干扰决策树对最优特征与最优分割点的选择,导致其分类性能受影响.

为此,本文提出一种基于随机投影[15]的决策树集成方法 Projection Forest(PJForest).随机投影是一种降维方法,该方法的基本原理是 Johnson-Lindenstrauss 引理(JL 引理)[16],JL 引理保证了当将高维空间的数据投影到低维空间时,能一定程度地保持数据的几何结构.由于随机投影具有十分优良的理论性质,使得它逐渐成为解决高维问题的一种重要方法,越来越多的学者将随机投影运用到高维问题的研究中.

Bingham 和 Mannila[17]利用随机投影对图像与文字数据进行降维,其实验结果表明随机投影能较好地保持数据信息,并且能较好地降低数据中的噪音,具有较低的计算复杂度;Dasgupta[18]则利用随机投影对高斯混合分布进行降维,其实验结果表明利用随机投影对高斯混合分布降维后,仍能较好地保持不同高斯分布之间的分离度;Wu 等[19]则利用随机投影与循环神经网络对文字进行分类,随机投影用来压缩从卷积神经网络学习出的高维特征,最后用循环神经网络作为分类器,实验结果表明计算时间显著降低的同时,分类精度几乎不变.这些实验结果均表明了随机投影能将高维数据嵌入一个低维空间,同时一定程度地保持数据原有信息,并且它能较好地降低数据中的噪音.

另外还有许多学者利用随机投影提升现有模型在高维情形下的表现.Durrant 和 Kaban[20]将随机投影应用到高维情形下线性判别分析分类器,其理论结果表明,随机投影对线性判别分析起了正则化的作用,克服了线性判别分析受维度灾难影响的缺点;Vinh 等[21]利用随机投影训练鲁棒的神经网络,其实验结果表明,随机投影能较好地对神经网络进行正则化,使得神经网络对带有噪音图像数据分类时有较稳定的分类性能.Cannings 和 Samworth[22]基于已有的研究成果,进一步研究了基分类器为 KNN,LDA,QDA 时随机投影集成,其理论结果表

明,随机投影能有效克服 KNN,LDA,QDA 遇到的维度灾难问题,并且利用集成能进一步提升分类器在高维情形下的表现.这些研究成果表明随机投影能作为一种正则化方式提升分类器在高维情形下的泛化能力.

本文提出基于随机投影的决策树集成方法 PJForest.该方法以决策树为基分类器,首先利用一系列随机投影对数据进行降维,然后基于降维后的数据构建相应的一系列决策树,最后通过多数投票法集成所有决策树,并产生预测结果.利用随机投影对数据降维,能保持数据的几何结构,并且比起一些已有的降维方法,比如 PCA,随机投影具有更低的计算复杂度,在变量维度 p 大于样本量 n 时,也能使用;更重要的是,随机投影通过对原始数据进行扰动,构建多棵好而不同的决策树,能丰富集成学习基分类器的多样性,通过将基于随机投影降维后的数据构建的一系列决策树集成,可得到具有较好泛化能力的 PJForest 分类器,而由于 PCA 总是将数据向最大方差方向进行投影,并不具备丰富集成学习的多样性的性质.本文通过模拟研究与实证分析测试 PJForest 的分类性能,并与已有方法进行比较,包括决策树、随机森林、极端随机森林[23]、Xgboost[24]、基于随机投影集成的 KNN 分类器[22].实验结果表明,PJForest 能较好对包含大量噪音的高维数据进行分类.本文还研究了 PJForest 的泛化误差的极限性质,理论结果表明当决策树个数趋于无穷时,PJForest 在满足一定条件的情形下会收敛,并给出了收敛速度.

1 方法介绍与理论性质

1.1 随机投影

随机投影方法的基本原理是 JL 引理[16].设 p 维空间的 n 个点$\{x_1,\cdots,x_n\}$,对 $\forall \epsilon\in(0,1)$,设整数 d 满足 $d>O(\frac{\ln(n)}{\epsilon^2})$,那么存在一个线性映射 $A:\mathrm{R}^p\to\mathrm{R}^d$,使得对所有 $i,j\in\{1,\cdots,n\}$,下式成立:

$$(1-\epsilon)\|x_i-x_j\|^2\leqslant\|Ax_i-Ax_j\|^2\leqslant(1+\epsilon)\|x_i-x_j\|^2.$$

JL 引理保证了对于存在于 p 维空间的数据可以嵌入一个 d 维空间,使得在一个任意小的常数因子 ϵ 下,数据两两之间的相对距离能够被保持.

满足 JL 引理的随机投影矩阵 $A_{d\times p}$ 有许多种构造方法,最常使用的是高斯随机投影矩阵[15],$A_{d\times p}$ 中的每个元素均服从高斯正态分布 $N(0,1)$.Li 等[25]提出还可以使用非常稀疏的随机投影矩阵,矩阵中的每个元素服从下述分布:

$$a_{ij}=\begin{cases}+\sqrt{s}\ \text{with probability}\ \dfrac{1}{2s};\\ 0\ \text{with probability}\ 1-\dfrac{1}{s};\\ -\sqrt{s}\ \text{with probability}\ \dfrac{1}{2s}.\end{cases}$$

$s=\sqrt{p}$ 或者可以更大,s 越大随机投影矩阵就越稀疏;这样构造的稀疏随机投影矩阵同样满足 JL 引理.比起高斯随机投影矩阵,稀疏随机投影矩阵的计算速度更快.

在实际应用中,除了考虑计算复杂度,还需要依据数据的特定结构选择具有特定分布的随机投影矩阵,比如当处理包含大量噪音的高维数据时,具有稀疏性的随机投影矩阵会比高斯随机投影矩阵更适合.在本文的模拟研究与实证分析中,均使用稀疏随机投影矩阵.

考虑一个简单的例子[1],利用稀疏随机投影矩阵对满足高斯分布的数据进行降维.如图 1 所示,20 维的数据被投影到 2 维平面,仍然一定程度保持着原有的几何结构,各类别之间具有相当的分离度;基于适当随机投影降维后的数据构造决策树时,分类准确率不会降低.并且,从图 1(b)~(d)可以看到,利用不同随机投影对数据进行降维,低维数据的结构是不同的,以此构建具有随机多样性特性的一系列决策树,进而可通过集成学习的方式提升集成分类器的泛化能力.

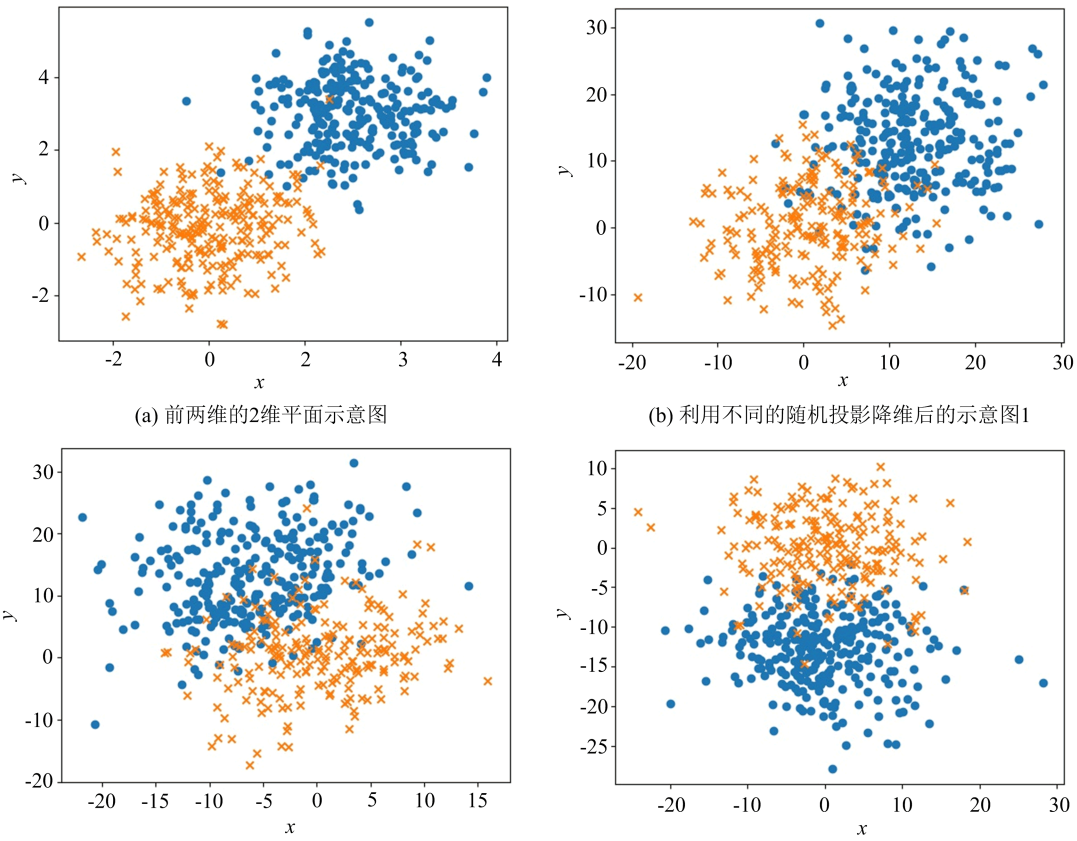
1.2 基于随机投影的决策树集成方法 PJForest

设(X,Y)表示定义在 $\mathrm{R}^p\times\{0,1\}$上的随机向量,具有联合分布 P;设训练集 $D_n=\{(x_1,y_1),\cdots,(x_n,y_n)\}$.设 $m(x)$表示 CART 决策树[5],有

$$m(x)=\sum_{i=1}^{t}c_iI(x\in V_i),$$

其中,$V_1,V_2,\cdots,V_t$ 表示决策树在特征空间分割出的 t 个互不相交的超立方体,c_i 表示 V_i 中多数样本所属的类别.为表明决策树与训练集 D_n 的关系,设基于训练集 D_n 构建的决策树为

$$m_n(X):=m(X;D_n).$$



(a) 前两维的2维平面示意图　(b) 利用不同的随机投影降维后的示意图1

(c) 利用不同的随机投影降维后的示意图2　(d) 利用不同的随机投影降维后的示意图3

图 1　$p=20$ 时的高斯分布，响应变量只与前两维有关

Fig. 1　Different two-dimensional projections of $p=20$ Gaussian observations

PJForest 通过随机投影对数据进行降维，对基于降维后的数据构建基分类器决策树. 设随机投影矩阵为 $A_{d\times p}$，通过如下方法将数据投影到低维空间：

$$\widetilde{x}=Ax\in \mathrm{R}^d,$$

则投影后的训练集记为 $D_n^A=\{(Ax_1,y_1),\cdots,(Ax_n,y_n)\}$. 设基于训练集 D_n^A 构建的决策树为 $m_n^A(X):=m(X;D_n^A)$. 将所有基于降维后的数据构建的决策树集成，就可以得到 PJForest 分类器.

对 $\forall B\in \mathrm{N}^+$，记 $A_1,A_2,\cdots,A_B$ 是满足 JL 引理的独立同分布的随机投影矩阵，且与 (X,Y) 独立. 设

$$M_n^{(B)}(x)=\frac{1}{B}\sum_{b=1}^{B}m_n^{A_b}(x),$$

那么 PJForest 分类器按如下定义：

$$C_n(x)=\begin{cases}1, M_n^{(B)}(x)\geqslant\dfrac{1}{2};\\ 0, M_n^{(B)}(x)<\dfrac{1}{2}.\end{cases}$$

其中，下标 n 表示训练集的样本量，A_b 表示第 b 个随机投影矩阵，B 表示基分类器的个数.

PJForest 与随机森林一个主要的不同点是随机森林中具有两个随机性扰动，Boostrap 抽样与随机选取特征变量；而 PJForest 的随机性扰动来自于随机投影. 比起基于随机投影的 KNN 集成分类器，PJForest 是以决策树为基分类器，由于决策树具有不稳定的特点，更能充分利用随机投影的随机性扰动去丰富集成学习的多样性.

1.3　极限性质

PJForest 是一种集成学习方法，而在集成学习中基分类器的个数对集成分类器的泛化能力有十分重要的影响. 本节主要研究当 B 趋向于无穷时，PJForest 的泛化误差的极限性质.

设 π_1 表示 $P(Y=1)$，π_0 表示 $P(Y=0)$；设已知 Y 情形下 X 的条件分布为

$$\mu_l(X)=P(X\mid Y=l), l\in\{0,1\};$$

设 $L(u(x),y)$ 为凸损失函数，则分类器 $u(x)$ 对应的泛化误差为

$$R(u)=\int_{R^p\times\{0,1\}}L(u(x),y)\mathrm{d}P(x,y).$$

为了简单起见，本文选取 0-1 损失函数进行讨论，即 $L(u(x),y)=I(u(x)\neq y)$，其中分类器 $u(x)$取值 0 或 1，所得结果不难推广至一般的凸损失函数的情形下.

给定数据 D_n，决策树 $m_n^{A_1}(X)$，$m_n^{A_2}(X)$…是独立同分布的随机函数. 设 E_A 表示对于随机投影取期望；设 $M_n^*(X)=E_A[m_n^A(X)]$，令

$$C_n^*(X)=I(M_n^*(X)\geqslant\frac{1}{2}),$$

$C_n^*(X)$可以视为决策树个数等于无穷情形下的 PJForest，记相应的泛化误差为

$$R(C_n^*)=\int_{R^p\times\{0,1\}}I(C_n^*(x)\neq y)\mathrm{d}P(x,y)=$$

$$\int_{R^p\times\{0,1\}}I(I(M_n^*(x)\geqslant\frac{1}{2})\neq y)\mathrm{d}P(x,y),$$

那么决策树个数为 B 的 PJForest 的泛化误差为

$$R(C_n)=$$

$$\int_{R^p\times\{0,1\}}I(I(M_n^{(B)}(x)\geqslant\frac{1}{2})\neq y)\mathrm{d}P(x,y).$$

$R(C_n)$是关于随机投影的随机变量，需进一步考虑 $E_A[R(C_n)]$的性质.

条件 1.1 对 $\forall l\in\{0,1\}$，条件分布函数 $F_l(s)=P(M_n^*(X)\leqslant s|Y=l)$ 有定义于区间$[0,1]$上的 Lebesgue 可测的概率密度函数 $f_l(s)$，并且 f_l 在区间$[0,1]$上有界，在$\frac{1}{2}$处可微，且在$\frac{1}{2}$邻域内满足 Lipschitz 条件.

说明 1.1 称 f_l 在$\frac{1}{2}$的邻域满足 Lipschitz 条件，如果存在常数 $\kappa>0$ 以及 $\delta>0$，使得 $\forall s,s'\in[\frac{1}{2}-\delta,\frac{1}{2}+\delta]$，满足$|f_l(s)-f_l(s')|\leqslant\kappa|s-s'|$. 这实际是要求 f_l 具备一定的光滑性.

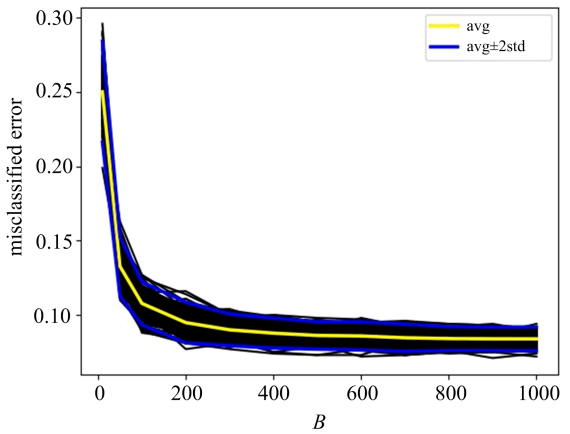
下面给出本文的主要定理.

定理 1.1 假设条件 1.1 满足，设 $R(C_n)$表示 $C_n(X)$的泛化误差，$R(C_n^*)$表示 $C_n^*(X)$的泛化误差，那么当 $B\to\infty$时下式成立：

$$E_A[R(C_n)]-R(C_n^*)=$$

$$\frac{\pi_0 f_0'(\frac{1}{2})+\pi_1 f'_1(\frac{1}{2})}{8B}+o(\frac{1}{B}).$$

定理 1.1 的证明参见附录. 定理 1.1 保证了当 B 趋于无穷时，PJForest 的泛化误差会收敛到 $R(C_n^*)$，并且收敛速度是$\frac{1}{B}$. 训练集 D_n 产生自高斯混合分布(见模拟研究)，固定训练集 D_n，重复训练 200 次 PJForest. 图 2 展示了 $R(C_n)$随 B 的变化，可以看到随着 B 变大时，PJForest 的分类错误率逐渐稳定并收敛，方差变得越来越小，与定理 1.1 结果相符. 通过选择较大 B 值的训练模型，可以得到一个分类性能较好的 PJForest 分类器.



黑线显示了 200 次 PJForest 分类错误率随 B 的变化，黄线代表分类错误率的平均值($E_A[R(C_n)]$的估计)，蓝线代表分类错误率的平均值加减两倍标准差.

图 2 分类错误率随 B 的变化

Fig. 2 Variation of classification error rate with B

$R(C_n^*)$是总体意义下的泛化误差. 而 $R(C_n^*)$与降维维度 d 有关，需要对 d 进行选优使得 $R(C_n^*)$最小. 根据 JL 引理，在给定 ϵ 的条件下，d 应满足 $d>O(\frac{\ln(n)}{\epsilon^2})$，如果 d 过小，会导致数据的原有结构被破坏，进而 $R(C_n^*)$就会变大. 在实际中，可以使用留一交叉验证的方法或 K 折交叉验证对 d 进行选优，这里以留一交叉验证为例说明. 假设 $d\in\{1,2,\cdots,p\}$，那么

$$\hat{d}=\underset{d}{\operatorname{argmin}}\frac{1}{n}\sum_{i=1}^{n}I(C_{n,i}^d(x_i)\neq y_i).$$

其中，$C_{n,i}^d$ 表示 PJForest 是在去除掉第 i 个样本的训练集上构建的.

2 模拟研究与实证分析

本节通过模拟研究与实证分析来测试 PJForest 的分类性能. 与之比较的算法包括决策树(Decision Tree)、随机森林(Random Forest)、极端随机森林(Extra Forest)、Xgboost 以及基于随机投影集成的 KNN 分类器(RP-KNN).

这些方法大多需要进行参数调优，本文使用网格搜索进行参数调优. 如决策树的最大深度从{5，

10,15,20,25,30}中根据 5 折交叉验证进行选择;随机森林的分裂节点时的最大特征数,从$\{\frac{p}{4},\frac{2p}{4},\frac{3p}{4},p\}$中根据交叉验证进行选择;RP-KNN 的 k 值与降维维度 d 中依据 5 折交叉验证进行最优选择,其中设定 $k\in\{3,5,7,9,11\}$,$d\in\{\frac{p}{3},\frac{2p}{3},\frac{3p}{4}\}$;PJForest 的降维维度 d 依据交叉验证从$\{\frac{p}{3},\frac{2p}{3},\frac{3p}{4}\}$中选择. 上述方法基分类器个数均设为 1 000,设置较大的基分类器个数有助于集成分类器达到最优分类性能.

在模拟研究中,使用分类错误率作为评估指标,每种情形均重复 100 次,最后计算 100 次分类错误率的平均值与标准差.

2.1　模拟研究

本节用人工数据来评估 PJForest 的分类性能. 人工数据由 $p(p\geqslant 2)$维高斯混合分布模型生成,按如下条件分布产生不同类别的数据:

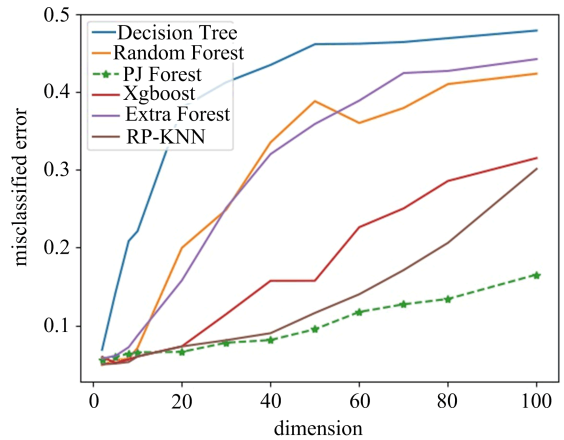
$$P(X\mid Y=r)\sim\frac{1}{2}N(\mu_r,\Sigma)+\frac{1}{2}N(-\mu_r,\Sigma).$$

其中,$r\in\{0,1\}$表示响应变量的类别;$\Sigma=I_{p\times p}$ 表示协方差矩阵;$\mu_1=(2,2,0,\cdots,0)_{p\times 1}$,$\mu_0=(2,-2,0,\cdots,0)_{p\times 1}$ 表示不同类别的 p 维均值向量. 从上式可以看出,响应变量 Y 只与前两维变量有关,其余 $p-2$ 维变量均为噪音.

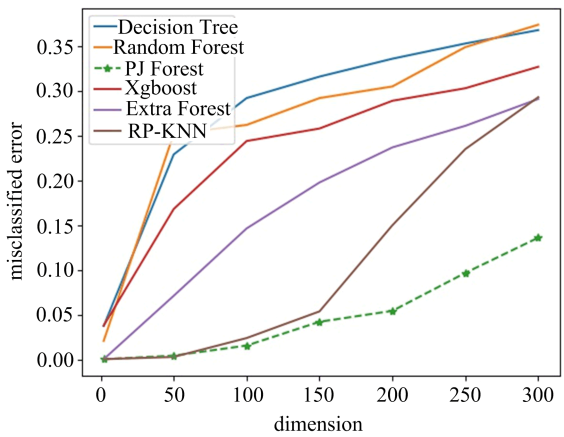
通过设置不同的参数对各分类器进行比较,第一组实验设置训练集样本量为 500,测试集样本量为 1 000, 令维度 p 从 2 增加到 100 时,比较各分类方法的分类错误率;第二组实验设置训练集样本量为 100,测试集样本量为 1 000, 令维度 p 从 2 增加到 300,比较各分类方法的分类错误率,在第二组实验中,按如下方式设置均值向量以增强信噪比:

$$\mu_0=(4.49,3.44,0,\cdots,0)_{p\times 1},$$
$$\mu_1=(3.44,-4.49,0,\cdots,0)_{p\times 1}.$$

第一组实验结果如图 3(a)所示. 从图 3(a)可以看到,当 p 从 2 增加到 100 时,决策树的分类错误率迅速上升;随机森林、Xgboost 与极端随机森林虽然比决策树表现好,但随着维度的增加,分类性能也受到较大影响;而 PJForest 与 RP-KNN 均取得了较好的表现;当 p 较高时,PJForest 在所有方法中取得了最好的分类表现,分类性能并没有受到较大



(a) 第一组实验



(b) 第二组实验

图 3　不同分类器分类错误率随维度 p 的变化

Fig. 3　Variation of classification error rate of different classifiers with p

影响,依然有十分出色与稳定的分类表现. 这是因为,当维度 p 较高时,数据存在大量噪音,随机森林在分裂节点选择最优特征时,受到影响选择了噪音特征,所以分类性能较差;而 PJForest 利用随机投影对样本进行降维,保持大量样本信息,经过集成能有效克服噪音的影响,所以有较好的分类表现.

在第二组实验中考虑特征维度 p 大于样本量的情形,结果如图 3(b)所示. 从图 3(b)可以看到,通过增强信噪比,所有方法在 p 较小时均取得了较好的分类表现,PJForest、极端随机森林、RP-KNN 几乎能完全正确的对数据进行分类;当 p 变大,超过样本量 100,直到 300 时,随机森林、决策树等方法的分类错误率也同时变大,当 $p=300$ 时,分类性能受到较大影响. RP-KNN 在 p 小于 200 时,有着较好的分类表现,而当 p 较大时,分类性能也变差;而 PJForest 在 p 从 100 增加到 300 时,均有较好的分类性能;并且比起其他已有的方法,依然具有最好的

分类表现.

这两组实验表明,PJForest 对高维数据中的噪音有较强的抵抗能力,与随机森林、Xgboost、极端随机森林等方法相比,有更好的分类性能.

2.2 实证分析

本节选取 3 个 UCI 公共数据集进行实证分析,这些数据集是信号数据或图像数据;信号数据与图像数据大多存在着大量噪音与冗余,并且具有较高的特征维度.本节通过这 3 个数据集对 PJForest 的分类性能进行测试.数据集的详细信息如下所述.

hill-valley data(noise 版本)来自 UCI 机器学习数据集,在二维平面中按特征顺序画出图像,那么这些图像看起来是一个山峰(hill)或者是山谷(valley),需要根据特征对山峰或山谷进行分类;hill-valley 训练集与测试集分别有 606 个样本,特征维度为 $p=100$. ESR(epileptic seizure recognition)来自 UCI 机器学习数据集,是生物信号数据,利用脑电波信号检测病人是否癫痫发作;特征维度 $p=178$,样本共包含 11 500 条记录. HAR(human activity recogition using smartphones)来自 UCI 机器学习数据集,是人类在表演不同动作时所佩戴的智能手机记录的信号数据,本文只考虑对动作上楼与下楼进行二分类;特征维度 $p=561$,训练样本包含 2 059 条记录,测试样本包含 891 条记录.

在实证分析中,划分训练集与测试集,对训练集抽取一定数量的样本训练模型,在测试集上评估分类错误率以及 F1-score,每种情形重复 100 次训练,得到结果是 100 次的平均值与标准差.比如 ESR 数据集,先划分训练集、测试集,再分别从训练集中抽取样本量为 100,200,500 的训练子集,在训练子集上训练 PJForest,在测试集上测试.

表 1 展示了各分类方法的分类错误率,可以看到,在所有方法中决策树的分类表现最差;极端随机森林在 ESR 数据集上取得了较好的分类表现;Xgboost 的分类表现一般,原因是 Xgboost 需要较大的样本量进行学习,当数据样本量不足时,Xgboost 易过拟合. PJForest 在 3 个数据集上均取得了最好的分类表现,表 1 的结果表明随机投影集成能有效提升决策树的分类性能.

表 1 各方法在公共数据集上的分类错误率

Tab. 1 The classification error rate of different methods on public datasets

公共数据集	样本量	分类方法					
		Decision Tree	Random Forest	Xgboost	PJForest	Extra Forest	RP-KNN
hill-valley	$n=100$	0.489(0.021)	0.490(0.024)	0.477(0.020)	**0.310(0.042)**	0.491(0.018)	0.431(0.032)
	$n=200$	0.483(0.017)	0.479(0.018)	0.459(0.019)	**0.223(0.041)**	0.477(0.018)	0.424(0.021)
	$n=500$	0.467(0.029)	0.452(0.014)	0.443(0.014)	**0.106(0.013)**	0.442(0.013)	0.401(0.010)
HAR	$n=100$	0.188(0.035)	0.099(0.021)	0.108(0.025)	**0.075(0.013)**	0.089(0.014)	0.102(0.024)
	$n=200$	0.177(0.028)	0.085(0.011)	0.093(0.022)	**0.068(0.009)**	0.079(0.010)	0.084(0.013)
	$n=500$	0.174(0.027)	0.074(0.007)	0.089(0.018)	**0.063(0.006)**	0.071(0.006)	0.080(0.007)
ESR	$n=100$	0.127(0.023)	0.075(0.167)	0.117(0.018)	0.065(0.012)	**0.061(0.008)**	0.066(0.0234)
	$n=200$	0.108(0.017)	0.054(0.006)	0.081(0.012)	**0.046(0.007)**	0.052(0.005)	0.048(0.0106)
	$n=500$	0.088(0.010)	0.045(0.004)	0.055(0.005)	**0.032(0.003)**	0.042(0.004)	0.037(0.004)

[注] 数据为 100 次实验的分类错误率的平均值,括号内是标准差.黑体数字表示最优的结果.

表 2 则比较了各分类方法的 F1-score,可以看到在 3 个数据集的不同样本量下,PJForest 均取得了最好的表现.PJForest 在 ESR 数据集 $n=100$ 时,分类错误率略大于极端随机森林,而 F1-score 却优于极端随机森林,这是因为 ESR 数据集具备一定的不均衡特性(癫痫病人数量少),而极端随机森林在预测时倾向于多数类样本,所以有较低的分类错误率,对少数类样本(癫痫病人)的识别较差,所以 F1-score 差于 PJForest.这个结果表明,PJForest 能一定程度地用于不均衡分类.

表 2　各方法在公共数据集上的 F1-score

Tab. 2　The F1-score of different methods on public datasets

公共数据集	样本量	分类方法					
		Decision Tree	Random Forest	Xgboost	PJForest	Extra Forest	RP-KNN
hill-valley	$n=100$	0.515(0.032)	0.518(0.038)	0.532(0.047)	**0.666(0.052)**	0.513(0.050)	0.486(0.096)
	$n=200$	0.522(0.021)	0.521(0.038)	0.533(0.032)	**0.752(0.038)**	0.522(0.036)	0.493(0.080)
	$n=500$	0.529(0.025)	0.548(0.018)	0.549(0.023)	**0.865(0.017)**	0.550(0.018)	0.561(0.051)
HAR	$n=100$	0.793(0.038)	0.892(0.019)	0.885(0.026)	**0.911(0.017)**	0.895(0.020)	0.910(0.019)
	$n=200$	0.801(0.030)	0.904(0.016)	0.904(0.023)	**0.922(0.009)**	0.912(0.013)	0.915(0.013)
	$n=500$	0.816(0.032)	0.915(0.008)	0.898(0.017)	**0.930(0.006)**	0.920(0.008)	0.923(0.005)
ESR	$n=100$	0.752(0.059)	0.788(0.057)	0.835(0.072)	**0.915(0.011)**	0.829(0.036)	0.803(0.059)
	$n=200$	0.809(0.05)	0.854(0.018)	0.874(0.03)	**0.937(0.006)**	0.862(0.014)	0.889(0.024)
	$n=500$	0.865(0.034)	0.886(0.008)	0.916(0.018)	**0.954(0.002)**	0.888(0.009)	0.918(0.010)

[注] 数据为 100 次实验的 F1-score 的平均值，括号内是标准差. 黑体数字表示最优的结果.

无论是以分类错误率还是以 F1-score 作为评价指标，PJForest 均有优异的表现. 实证分析的结果表明，对于图像、信号数据，往往维度较高且包含大量噪音，随机投影能有效地对这类数据降维，经过集成能克服噪音的影响，这使得 PJForest 比起随机森林、Xgboost 等方法有更好的分类表现.

3　结论

本文提出了一种基于随机投影的决策树集成分类方法 PJForest. 本文研究表明，一方面随机投影能有效对高维数据进行降维，保持数据几何结构的信息；另一方面随机投影还能丰富集成学习的多样性，并且经过集成可有效克服噪音的影响，这使得 PJForest 有较好的泛化性能. 对于 PJForest，本文证明了其泛化误差 $E_A[R(C_n)]$ 以 $\frac{1}{B}$ 的速度收敛到 $R(C_n^*)$；通过选取较大的 B 能得到具有较好分类性能的 PJForest 分类器；降维维度 d 可以使用留一交叉验证或 K 折交叉验证的方法进行优选. 本文还通过模拟研究与实证分析对 PJForest 的有效性进行测试. 模拟研究表明 PJForest 能有效地对包含有大量噪音的高维数据进行分类；实证分析表明 PJForest 十分适合处理高维有噪音的图像数据、信号数据，与已有的诸如随机森林、Xgboost 等方法相比，有更好的分类表现.

本文方法 PJForest 是在假设数据均衡条件下提出的，许多实际中的应用还要求处理严重不均衡数据，PJForest 可以进一步地拓展到高维不均衡数据的情形下. 一个简便的方法是在 PJForest 进行投票时，根据数据的先验分布去决定投票阈值 α，即令 PJForest 分类器为

$$C_n(x)=I(M_n^{(B)}(x)\geqslant\alpha).$$

在数据分布均衡的情形下，那么 $\alpha=\frac{1}{2}$；在数据非均衡情形下，使用数据驱动的方式决定投票阈值 α，即可以根据已有数据去估计类别的先验分布：

$$\alpha=\frac{\sum_{i=1}^{n}I(y_i=1)}{n}.$$

参考文献(References)

[1] FAN J, HAN F, LIU H. Challenges of big data analysis[J]. National Science Review, 2014, 1 (2): 293-314.

[2] FAN J, LI R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery[EB/OL]. [2019-03-01] https://arxiv.org/abs/math/0602133.

[3] FAN J, FAN Y. High dimensional classification using features annealed independence rules [J]. Annals of Statistics, 2008, 36(6): 2605-2637.

[4] DONOHO D L. High-dimensional data analysis: The curses and blessings of dimensionality [C]// Math Challenges of the 21st Century. Providence, RI: American Mathematical Society, 2000.

[5] BREIMAN L, FRIEDMAN J, STONE C J, et al.

Classification and Regression Trees[M]. New York: CRC Press, 1984.

[6] SAVE P, TIWAREKAR P, JAIN K N, et al. A novel idea for credit card fraud detection using decision tree [J]. International Journal of Computer Applications, 2017, 161(13): 6-9.

[7] GOKGOZ E, SUBASI A. Comparison of decision tree algorithms for EMG signal classification using DWT [J]. Biomedical Signal Processing and Control, 2015, 18: 138-144.

[8] SALMAM F Z, MADANI A, KISSI M. Facial expression recognition using decision trees [C]// 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV). IEEE, 2016, 1: 125-130.

[9] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24(2):123-140.

[10] HO T K. Random decision forests[C]// Proceedings of 3rd International Conference on Document Analysis and Recognition: Volume 1. IEEE, 1995: 278-282.

[11] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[12] DIAZ-URIARTE R, DE ANDRES S A. Gene selection and classification of microarray data using random forest[J]. BMC Bioinformatics, 2006, 7: 3.

[13] JOELSSON S R, BENEDIKTSSON J A, SVEINSSON J R. Random forest classifiers for hyperspectral data [C]// Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. IEEE, 2005.

[14] XU B, HUANG J Z, WILLIAMS G, et al. Classifying very high-dimensional data with random forests built from small subspaces [J]. International Journal of Data Warehousing and Mining (IJDWM), 2012, 8(2): 44-63.

[15] DASGUPTA S, GUPTA A. An elementary proof of the Johnson-Lindenstrauss lemma [J]. Random Structures & Algorithms, 1999, 22(1): 1-5.

[16] JOHNSON W B, LINDENSTRAUSS J. Extensions of Lipschitz mappings into a Hilbert space [J]. Contemporary Mathematics, 1984, 26: 189-206.

[17] BINGHAM E, MANNILA H. Random projection in dimensionalityreduction: Applications to image and text data [C]// Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2001: 245-250.

[18] DASGUPTA S. Experiments with random projection [EB/OL]. [2019-03-01] https://arxiv.org/abs/1301.3849.

[19] WU R, YANG S, LENG D, et al. Random projected convolutional feature for scene text recognition[C]// 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2016: 132-137.

[20] DURRANT R J, KABAN A. Random projections as regularizers: Learning a linear discriminant from fewer observations than dimensions[J]. Machine Learning, 2015, 99(2): 257-286.

[21] VINH N X, ERFANI S, PAISITKRIANGKRAI S, et al. Training robust models using random projection [C]// 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016: 531-536.

[22] CANNINGS T I, SAMWORTH R J. Random-projection ensemble classification[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2017, 79(4): 959-1035.

[23] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

[24] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[25] LI P, HASTIE T J, CHURCH K W. Very sparse random projections[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006: 287-296.

[26] LOPES M E. A sharp bound on the computation-accuracy tradeoff for majority voting ensembles[EB/OL]. [2019-03-01] https://arxiv.org/abs/1303.0727v2.

[27] BOUCHERON S, LUGOSI G, MASSART P. Concentration Inequalities: A Nonasymptotic Theory of Independence[M]. Oxford, UK: Oxford University Press, 2013.

[28] VAN DER VAART A W. Asymptotic Statistics: Volume 3 [M]. Cambridge, UK: Cambridge University Press, 2000.

附录

定理 1.1 的证明与文献[26]类似,本文将其拓展到 PJForest 情形下.

定理 1.1 证明　首先推导一个分解公式,注意到

$$R(C_n)=\int_{R^p\times\{0,1\}} I(C_n(x)\neq y)\mathrm{d}P(x,y)=\pi_1\int_{R^p} I(C_n(x)\neq 1)\mathrm{d}\mu_1(x)+\pi_0\int_{R^p} I(C_n(x)\neq 0)\mathrm{d}\mu_0(x).$$

定义

$$\mathrm{ERR}_{B,0}=\int_{R^p} I(C_n(x)\neq 0)\mathrm{d}\mu_0(x)=\int_{R^p} I(M_n^{(B)}(x)\geqslant\frac{1}{2})\mathrm{d}\mu_0(x)\ ,$$

$$\mathrm{ERR}_{B,1}=\int_{R^p} I(C_n(x)\neq 1)\mathrm{d}\mu_1(x)=\int_{R^p} I(M_n^{(B)}(x)\leqslant\frac{1}{2})\mathrm{d}\mu_1(x).$$

那么有如下分解公式成立:

$$R(C_n)=\pi_0\mathrm{ERR}_{B,0}+\pi_1\mathrm{ERR}_{B,1}\tag{A1}$$

式中,$\mathrm{ERR}_{B,l}$ 表示在 $Y=l$ 时的泛化误差. 于是只需分别考虑 $\mathrm{ERR}_{B,0}$,$\mathrm{ERR}_{B,1}$. 由于 $\mathrm{ERR}_{B,0}$ 与 $\mathrm{ERR}_{B,1}$ 是类似的,本文只考虑 $\mathrm{ERR}_{B,1}$ 的情形.

下面证明 $E_A[\mathrm{ERR}_{B,1}]$的收敛.

定义 $H_B(x)=I(M_n^{(B)}(x)\leqslant\frac{1}{2})$,那么可以将 $\mathrm{ERR}_{B,1}$ 写成

$$\mathrm{ERR}_{B,1}=\int_{R^p} I(M_n^{(B)}(x)\leqslant\frac{1}{2})\mathrm{d}\mu_1(x)=\int_{R^p} H_B(x)\mathrm{d}\mu_1(x),$$

又设 $h_B(x)=E_A[H_B(x)]=P_A(M_n^{(B)}(x)\leqslant\frac{1}{2})$,那么 $E_A[\mathrm{ERR}_{B,1}]=\int_{R^p} h_B(x)\mathrm{d}\mu_1(x)$. 注意到 $M_n^*(x)$与 $h_B(x)$的联系,设 $U_1,\cdots,U_B\sim U[0,1]$,令 $g_B(s)=P(\frac{1}{B}\sum_{i=1}^{B} I(U_i\leqslant s)\leqslant\frac{1}{2})$,因为随机变量序列$\{M_n^{A_i}(x)\}$与$\{I(U_i\leqslant M_n^*(x))\}$都服从 Bernoulli$(M_n^*(x))$ 分布.所以对所有的 x 与 B 有下式成立:

$$h_B(x)=g_B(M_n^*(x))\tag{A2}$$

于是可以利用条件 1.1 与式(A2)作变量替换,有

$$E_A[\mathrm{ERR}_{B,1}]=\int_{R^p} h_B(x)\mathrm{d}\mu_1(x)=E_{X|Y=1}[h_B(x)]=E_{X|Y=1}[g_B(M_n^*(x))]=$$

$$E_{M_n^*(X)|Y=1}[g_B(x)]=\int_0^1 g_B(\theta)f_1(\theta)\mathrm{d}\theta=\int_0^{\frac{1}{2}}(g_B(\theta)f_1(\theta)+g_B(1-\theta)f_1(1-\theta))\mathrm{d}\theta,$$

最后一个等式是在积分区间$[\frac{1}{2},1]$中使用 $1-\theta$ 代替 θ 而得. 又 $F_1(\frac{1}{2})=\int_0^{\frac{1}{2}} f_1(\theta)\mathrm{d}\theta$,则有

$$E_A[\mathrm{ERR}_{B,1}]-F_1(\frac{1}{2})=\int_0^{\frac{1}{2}}(g_B(\theta)-1)f_1(\theta)+g_B(1-\theta)f_1(1-\theta)\mathrm{d}\theta.$$

利用 g_B 的对称性,

$$g_B(1-\theta)=P(\frac{1}{B}\sum_{i=1}^{B} I(U_i\leqslant 1-\theta)\leqslant\frac{1}{2})=P(\frac{1}{B}\sum_{i=1}^{B} I(U_i\geqslant\theta)\leqslant\frac{1}{2})=$$

$$1-P(\frac{1}{B}\sum_{i=1}^{B} I(U_i\leqslant\theta)\leqslant\frac{1}{2})=1-g_B(\theta),$$

可得 $E_A[\mathrm{ERR}_{B,1}]-F_1(\frac{1}{2})=\int_0^{\frac{1}{2}} g_B(1-\theta)(f_1(1-\theta)-f_1(\theta))\mathrm{d}\theta$. 作变量替换:$\theta=\frac{1}{2}-\frac{u}{\sqrt{B}}$,$u\in[0,\frac{\sqrt{B}}{2}]$, 那么

$$B(E_A[\mathrm{ERR}_{B,1}]-F_1(\frac{1}{2}))=\int_0^{\frac{\sqrt{B}}{2}} g_B(\frac{1}{2}+\frac{u}{\sqrt{B}})\frac{f_1(\frac{1}{2}+\frac{u}{\sqrt{B}})-f_1(\frac{1}{2}-\frac{u}{\sqrt{B}})}{\frac{2u}{\sqrt{B}}}2u\,\mathrm{d}u \tag{A3}$$

记式(A3)积分号内的式子为 $\psi_B(u)$. 需要利用控制收敛定理证明式(A3)的收敛.

首先证 $\psi_B(u)$有界. 因为 f_1 在$\frac{1}{2}$的邻域内满足 Lipschitz 条件,所以$\exists\kappa>0$,对充分大的 B,下式成立:

$$\left|\frac{f_1(\frac{1}{2}+\frac{u}{\sqrt{B}})-f_1(\frac{1}{2}-\frac{u}{\sqrt{B}})}{\frac{2u}{\sqrt{B}}}\right|\leqslant\kappa.$$

对于 g_B,利用 Hoeffding's 不等式等[27], 可得 $g_B(\frac{1}{2}+\frac{u}{\sqrt{B}})\leqslant \mathrm{e}^{-2u^2}$, 其中 $0\leqslant u\leqslant\frac{\sqrt{B}}{2}$,所以可证有界性:

$$B(E_A[\mathrm{ERR}_{B,1}]-F_1(\frac{1}{2}))=\int_0^{\frac{\sqrt{B}}{2}}\psi_B(u)\mathrm{d}u\leqslant\int_0^{\frac{\sqrt{B}}{2}}2\kappa u\,\mathrm{e}^{-2u^2}\mathrm{d}u=\frac{\kappa}{2}(1-\mathrm{e}^{-\frac{B}{2}})\to\frac{\kappa}{2}<\infty.$$

接着,只需证 $B(E_A[\mathrm{ERR}_{B,1}]-F_1(\frac{1}{2}))$的收敛. 下式是显而易见的:

$$\frac{f_1(\frac{1}{2}+\frac{u}{\sqrt{B}})-f_1(\frac{1}{2}-\frac{u}{\sqrt{B}})}{\frac{2u}{\sqrt{B}}}\to f_1'(\frac{1}{2}).$$

故只需要考虑 $g_B(\frac{1}{2}+\frac{u}{\sqrt{B}})$. 这里利用经验过程的方法去证明所要求的收敛,设 $F_t(\theta)=\frac{1}{B}\sum_{i=1}^{B}I(U_i\leqslant\theta)$, 那么有

$$g_B(\frac{1}{2}+\frac{u}{\sqrt{B}})=P(F_B(\frac{1}{2}+\frac{u}{\sqrt{B}})\leqslant\frac{1}{2})=P(\sqrt{B}(F_B(\frac{1}{2}+\frac{u}{\sqrt{B}})-(\frac{1}{2}+\frac{u}{\sqrt{B}}))\leqslant-u),$$

由 Donsker's theorem[28],如果 序列$\{\theta_B\}\in[0,1]$ 收敛到 θ_0, 那么下式弱收敛成立:

$$\sqrt{B}(F_B(\theta_B)-\theta_B)\to B_{\theta_0}.$$

其中,B 是标准的布朗桥, $B_{\theta_0}\sim N(0,\theta_0(1-\theta_0))$.

又$\frac{1}{2}+\frac{u}{\sqrt{B}}\to\frac{1}{2}$,故 $g_B(\frac{1}{2}+\frac{u}{\sqrt{B}})\to\Phi(-2u)$, 那么由控制收敛定理立即可证

$$B(E_A[\mathrm{ERR}_{B,1}]-F_1(\frac{1}{2}))\to 2f_1'(\frac{1}{2})\int_0^{\infty}u\Phi(-2u))\mathrm{d}u=\frac{1}{8}f_1'(\frac{1}{2}).$$

从而可得

$$E_A[\mathrm{ERR}_{B,0}]=1-F_0(\frac{1}{2})+\frac{1}{8B}f_0'(\frac{1}{2})+o(\frac{1}{B}),$$

$$E_A[\mathrm{ERR}_{B,1}]=F_1(\frac{1}{2})+\frac{1}{8B}f_1'(\frac{1}{2})+o(\frac{1}{B}).$$

由分解公式(A1)容易证明 $E_A[R(C_n)]$的收敛:

$$E_A[R(C_n)]=\pi_0E_A[\mathrm{ERR}_{B,0}]+\pi_1E_A[\mathrm{ERR}_{B,1}]=$$

$$\pi_0(1-F_0(\frac{1}{2})+\frac{1}{8B}f_0'(\frac{1}{2})+o(\frac{1}{B}))+\pi_1(F_1(\frac{1}{2})+\frac{1}{8B}f_1'(\frac{1}{2})+o(\frac{1}{B}))=$$

$$R(C_n^*)+\frac{\pi_0f_0'(\frac{1}{2})+\pi_1f_1'(\frac{1}{2})}{8B}+o(\frac{1}{B}).$$