Random-projection ensemble classification 论文阅读与思考

原创 Shian150629 2021-02-12 21:41:59 34 收藏 编辑 版权

分类专栏: 论文阅读 论文阅读与复现 文章标签: 机器学习

论文阅读 同时被 2 个专栏收录 0 订阅 3 篇文章

The Email address begins with shianlin2084.

随机投影集成分类

Shian150629

4.有一点没想明白，如何知道是测试集上最好的?

1. 论文泛读

1.1. 标题

随机投影集成分类

1.2. 摘要

1.2.1. 方案

这是一个通用方案：对高维数据分类，使用随机投影，将特征向量降维至低维空间，然后使用任意基分类器，选出合适的进行结合。将随机矩阵划分成不相关的组。在每个组里选出服从最小**测试误差**（？）的估计。然后集成这些结果，使用数据驱动型投票阈值决定最终结果。我们的理论结果阐明了增加投影数量对性能的影响。

1.2.2. 效果

1.我们的理论结果阐明了增加投影数量对性能的影响
2.此外，在充分降维假设所隐含的边界条件下，我们证明了随机投影集成分类器的测试超额风险可以由不依赖于原始数据维数的项来控制
3.随着预测数量的增加，一个项变得可以忽略不计
4.通过大量的模拟研究，将该分类器与其他几种常用的高维分类器进行了实证比较，显示出其优异的有限样本性能

1.3. 介绍

1.3.1. 前人方法

- 符号：p：矩阵维度；n（训练）样本数目
- LDA等一系列不适应高维
- 使用特征提取；也有使用软阈值获取稀疏边界
- 使用正则项

1.3.2. RPEnsemble

- 随机投影：the celebrated Johnson–Lindenstrauss Lemma

> 2002). This lemma states that, given $x_1, \ldots, x_n \in \mathbb{R}^p$, $\epsilon \in (0,1)$ and $d > \frac{8 \log n}{\epsilon^2}$, there exists a linear map $f : \mathbb{R}^p \to \mathbb{R}^d$ such that
>
> $$(1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2,$$

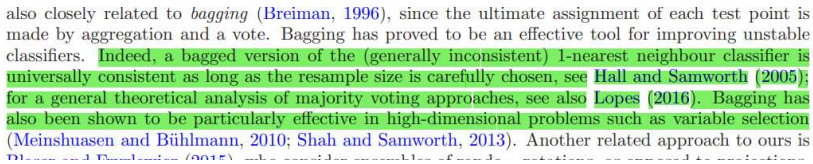
注意，在这个保证了对偶距离的函数f，可以使用哈尔测度上的随机投影分布，在随机多项式时间里找到【第三章证明，但是实际操作可以见公布的代码】。有趣的是，在该引理中的投影下界d，并不依赖于初始维度p。这个下界，常数因子是最优的 ===> 随机矩阵投影可以大量节约时间。当维度p大于log n的时候，使用随机矩阵投影可以有与原来矩阵具有相同甚至更好的统计意义上的表现。【这句后面有讲，只用一个随机投影矩阵的例子】

- 集成：

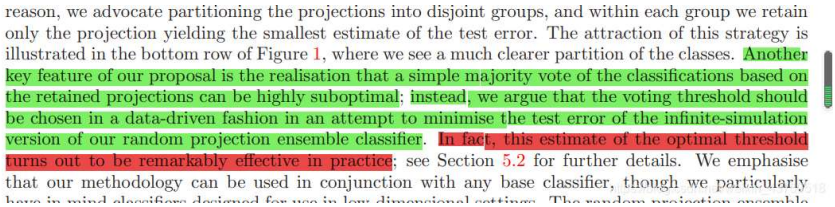
> instance, Marzetta, Tucci and Simon (2011) considered estimating a $p \times p$ population inverse covariance (precision) matrix using $B^{-1} \sum_{b=1}^{B} \mathbf{A}_b^T (\mathbf{A}_b \hat{\Sigma} \mathbf{A}_b^T)^{-1} \mathbf{A}_b$, where $\hat{\Sigma}$ denotes the sample covariance matrix and $\mathbf{A}_1, \ldots, \mathbf{A}_B$ are random projections from $\mathbb{R}^p$ to $\mathbb{R}^d$. Lopes, Jacob and Wainwright (2011) used this estimate when testing for a difference between two Gaussian population means in high dimensions, while Durrant and Kabán (2015) applied the same technique in Fisher's linear discriminant for a high-dimensional classification problem.

- 大概意思是说，第一个2011代指的那篇论文，用那个公式估计一个p*p大小的总体逆协方差。后面两个引用都用这篇论文的idea进行实验。这个地方，请注意，并不是限定为仅可以在分类上使用的

- bagging:


also closely related to *bagging* (Breiman, 1996), since the ultimate assignment of each test point is made by aggregation and a vote. Bagging has proved to be an effective tool for improving unstable classifiers. Indeed, a bagged version of the (generally inconsistent) 1-nearest neighbour classifier is universally consistent as long as the resample size is carefully chosen, see Hall and Samworth (2005); for a general theoretical analysis of majority voting approaches, see also Lopes (2016). Bagging has also been shown to be particularly effective in high-dimensional problems such as variable selection (Meinshuasen and Bühlmann, 2010; Shah and Samworth, 2013). Another related approach to ours is

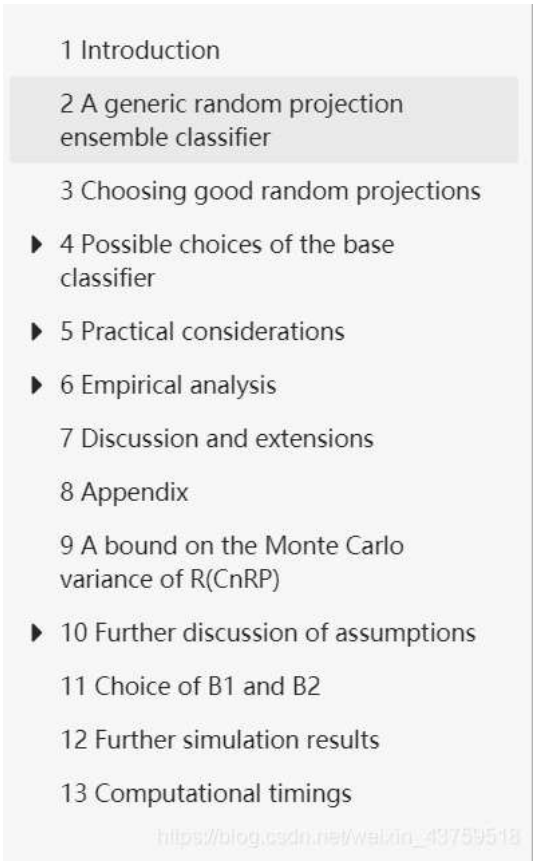
- 筛选数据的理论支持


reason, we advocate partitioning the projections into disjoint groups, and within each group we retain only the projection yielding the smallest estimate of the test error. The attraction of this strategy is illustrated in the bottom row of Figure 1, where we see a much clearer partition of the classes. Another key feature of our proposal is the realisation that a simple majority vote of the classifications based on the retained projections can be highly suboptimal: instead, we argue that the voting threshold should be chosen in a data-driven fashion in an attempt to minimise the test error of the infinite-simulation version of our random projection ensemble classifier. In fact, this estimate of the optimal threshold turns out to be remarkably effective in practice; see Section 5.2 for further details. We emphasise that our methodology can be used in conjunction with any base classifier, though we particularly

这段就说怎么筛选，怎么确定阈值的

- 论文结构：
  略。跳过理论部分

1.4. 小标题



通读可以发现，理论部分优先跳过，其他泛读，结合公布的代码与实验进行理解与模仿。第八节后都是附录，选择参数的要看

1.5. 结论和讨论

1. 这玩意儿是个框架，啥分类器都可以往里面套
2. 可以给投票加权重

Shian150629

3. 面对多分类问题的拓展：

Many practical classification problems involve $K > 2$ classes. The main issue in extending our methodology to such settings is the definition of C_n^{RP} analogous to (2). To outline one approach, let

$$\nu_{n,r}(x) := \frac{1}{B_1}\sum_{b_1=1}^{B_1} \mathbb{1}_{\{C_n^{\mathbf{A}_{b_1}}(x)=r\}}$$

for $r = 0, 1, \ldots, K-1$. Given $\alpha_0, \ldots, \alpha_{K-1} > 0$ with $\sum_{r=0}^{K-1} \alpha_r = 1$, we can then define

$$C_n^{\mathrm{RP}}(x) := \underset{r=0,\ldots,K-1}{\operatorname{sargmax}} \{\alpha_r \nu_{n,r}(x)\},$$

4. 面对其他随机投影的选择

- **维度过高**，例如是上千级别的，生成随机投影就很花时间，这个时候，就有

random projections to explore adequately the space $\mathcal{A}_{d\times p}$. As a mathematical quantification of this, the cardinality of an ϵ-net in the Euclidean norm of the surface of the Euclidean ball in $\mathbb{R}^p$ increases exponentially in p (e.g. Vershynin, 2012). In such challenging problems, one might restrict the projections $\mathbf{A}$ to be axis-aligned, so that each row of $\mathbf{A}$ consists of a single non-zero component, equal to 1, and $p-1$ zero components. There are then only $\binom{p}{d} \leq p^d/d!$ choices for the projections, and if d is small, it may be feasible even to carry out an exhaustive search. Of course, this approach loses one

也就是使用一种矩阵A。A的每一行都只有一个非零元素（是1）。请注意A是d*p大小的。当然，这种方法丢失了RPEnsemble里面最有吸引力的地方（这个地方就是它与正交变换是等价的）。要证明相应的理论是可以的，但这种情况下，要获得良好的分类，就不可避免地需要更大的结构。【RNM】

5. 虽然这玩意解释性不好，但选出来的随机投影暗含的权重表明了不同变量的相对重要性。也可以从这一方向来理解：RPEnsemble分类器生成了变量排序
6. 类似于分抽样和自举抽样，我们可以认为对原始数据的每一个随机投影，以及在许多不同扰动下观察到的效果，往往是统计学家所寻求的"稳定"效果
7. 为啥RPEnsemble对分类问题有吸引力?
   a. 因为它们能够从数据中识别出"好的"随机预测
   b. 我们可以从选定的预测中汇总结果
   预计这两个以上特性将在确定相关方法的未来应用领域中发挥重要作用

1.6. 图表

- 分类器使用LDA,QDA,KNN
- fig 1是用200例50维的数据，分别随便投影和精挑细选（上下）成为2维，LDA,QDA,KNN（左右）来证明下面的比上面的好
- fig2 的黑线是平均误差，上下两个标准差（红线），在超过20组B_1,B_2上得到的。使用模型是model 2，其他参数是n = 50,p = 100,d = 5,B_2都是50。三张图是三个分类器
- fig3 中，变动的是样本数量KaTeX parse error: Undefined control sequence: \n at position 1: \n和π_i。固定的分类器是QDA，样本维度$p = 100$，$d = 2$红线是估计，黑线是真实值。使用模型3
- fig4，5 是直方图，第10节的，先跳过；发现都是附录的内容，跳
- 实验的话是4个数值实验，$n = 50, 200, 1000$，$p = 100, 1000$。有两种不同的先验概率。**使用高斯投影**，令$B_1 = 500, B_2 = 50$。表格1和2是风险估计和标准差，$p = 100$，$\pi_1 = 0.5$，$n_{test} = 1000$是测试集大小，$l = 1, .., N_{reps}$，$N_{reps} = 100$。也就是把这些实验重复100次。计算出每次在**测试集**上的情况来取平均值，有

that is classified incorrectly in the rth repeat of the experiment. The overall risk estimate presented is $\widehat{\text{Risk}} := \frac{1}{N_{\text{reps}}} \sum_{l=1}^{N_{\text{reps}}} \hat{R}_l$. Note that

$$\mathbb{E}\{\widehat{\text{Risk}}\} = \mathbb{E}\{R(C_n^{\text{RP}})\}$$

and

$$\begin{aligned}\text{Var}(\widehat{\text{Risk}}) &= \frac{1}{N_{\text{reps}}} \text{Var}(\hat{R}_1) \\ &= \frac{1}{N_{\text{reps}}} \left[\mathbb{E}\left\{ \frac{\mathbf{E}\{R(C_n^{\text{RP}})\}[1 - \mathbf{E}\{R(C_n^{\text{RP}})\}]}{n_{\text{test}}} \right\} + \text{Var}\left[\mathbf{E}\{R(C_n^{\text{RP}})\} \right] \right].\end{aligned}$$

We therefore estimate the standard error in the tables below by

$$\hat{\sigma} := \frac{1}{N_{\text{reps}}^{1/2}} \left\{ \frac{\widehat{\text{Risk}}(1 - \widehat{\text{Risk}})}{n_{\text{test}}} + \frac{n_{\text{test}} - 1}{n_{\text{test}} N_{\text{reps}}} \sum_{l=1}^{N_{\text{reps}}} (\hat{R}_l - \widehat{\text{Risk}})^2 \right\}^{1/2}.$$

加粗是最好的；我们还强调的是风险评估在一个最小标准误差内的方法
这个图片，介绍了风险估计的期望和方差情况，也推出了标准差估计的式子

- 模型1【稀疏类边界】

0.1.1. *Sparse class boundaries*

Model 1: Here, $X|\{Y=0\} \sim \frac{1}{2}N_p(\mu_0, \Sigma) + \frac{1}{2}N_p(-\mu_0, \Sigma)$, and $X|\{Y=1\} \sim \frac{1}{2}N_p(\mu_1, \Sigma) + \frac{1}{2}N_p(-\mu_1, \Sigma)$, where, for $p = 100$, we set $\Sigma = I_{100\times 100}$, $\mu_0 = (2, -2, 0, \ldots, 0)^T$ and $\mu_1 = (2, 2, 0, \ldots, 0)^T$.

- 模型2【旋转的稀疏正交】

Model 2: Here, $X|\{Y=0\} \sim N_p(\Omega_p\mu_0, \Omega_p\Sigma_0\Omega_p^T)$, and $X|\{Y=1\} \sim N_p(\Omega_p\mu_1, \Omega_p\Sigma_1\Omega_p^T)$, where Ω_p is a $p \times p$ rotation matrix that was sampled once according to Haar measure, and remained fixed thereafter, and we set $\mu_0 = (3, 3, 3, 0, \ldots, 0)^T$ and $\mu_1 = (0, \ldots, 0)^T$. Moreover, Σ_0 and Σ_1 are block diagonal, with blocks $\Sigma_r^{(1)}$, and $\Sigma_r^{(2)}$, for $r = 0, 1$, where $\Sigma_0^{(1)}$ is a 3×3 matrix with diagonal entries equal to 2 and off-diagonal entries equal to $1/2$, and $\Sigma_1^{(1)} = \Sigma_0^{(1)} - I_{3\times 3}$. In both classes $\Sigma_r^{(2)}$ is a $(p-3) \times (p-3)$ matrix, with diagonal entries equal to 1 and off-diagonal entries equal to $1/2$.

- 模型3【特征独立】

Model 3: Here, $P_0 = N_p(\mu, I_{p\times p})$, with $\mu = \frac{1}{\sqrt{p}}(1, \ldots, 1, 0, \ldots, 0)^T$, where μ has $p/2$ non-zero components, while P_1 is the distribution of p independent components, each with a standard Laplace distribution.

- 模型4【t分布特征】

Model 4: Here, $X|\{Y=r\} = \mu_r + \frac{Z_r}{\sqrt{U_r/\nu_r}}$, where $Z_r \sim N_p(0, \Sigma_r)$ independent of $U_r \sim \chi^2_{\nu_r}$, for $r = 0, 1$. That is, P_r is the multivariate t-distribution centred at μ_r, with ν_r degrees of freedom and shape parameter Σ_r. We set $\mu_0 = (1, \ldots, 1, 0, \ldots, 0)^T$, where μ_0 has 10 non-zero components, $\mu_1 = 0$, $\nu_0 = 2$, $\nu_1 = 1$, $\Sigma_0 = (\Sigma_{j,k})$, where $\Sigma_{j,j} = 1$, $\Sigma_{j,k} = 0.5$ if $\max(j,k) \le 10$ and $j \ne k$, $\Sigma_{j,k} = 0$ otherwise, and $\Sigma_1 = I_{p\times p}$.

- 表格1 里，这误差还不小啊尤其模型1
- 表格2 里，n小的时候明显都很大
- 真实数据模型的样本量和数据维度也不大

2. 论文精读

2.1. 算法框架与符号标记【对应第二节】

Finally in this section, we introduce the following general notation used throughout the paper. For a sufficiently smooth real-valued function g defined on a neighbourhood of $t \in \mathbb{R}$, let $\dot{g}(t)$ and $\ddot{g}(t)$ denote its first and second derivatives at t, and let $\lfloor t \rfloor$ and $[\![t]\!] := t - \lfloor t \rfloor$ denote the integer and fractional part of t respectively.

- (X, Y)是来自联合分布P。这个分布具有以下特征:

1. $\pi_1 := P(Y = 1)$

2. P_r描述条件概率$X|Y=r,\ r=0,1$
   因此，我们可以推出
3. $\pi_0 := P(Y=0) = 1-\pi_1$
4. 边际分布为P_X
5. $\eta(x) := P(Y=1|X=x)$
6. 所有的分类器【二分类】的集合

$\eta(x) := \mathbb{P}(Y = 1|X = x)$ for the regression function. Recall that a *classifier* on $\mathbb{R}^p$ is a Borel measurable function $C : \mathbb{R}^p \to \{0, 1\}$, with the interpretation that we assign a point $x \in \mathbb{R}^p$ to class $C(x)$. We let $\mathcal{C}_p$ denote the set of all such classifiers.

7. 分类器在测试集上的误差

The test error of a classifier C is‡

$$R(C) := \int_{\mathbb{R}^p \times \{0,1\}} \mathbb{1}_{\{C(x) \neq y\}}\, dP(x, y),$$

and is minimised by the *Bayes* classifier

$$C^{\text{Bayes}}(x) := \begin{cases} 1 & \text{if } \eta(x) \geq 1/2; \\ 0 & \text{otherwise} \end{cases}$$

(e.g. Devroye, Györfi and Lugosi, 1996, p. 10). Its risk is $R(C^{\text{Bayes}}) = \mathbb{E}[\min\{\eta(X), 1-\eta(X)\}]$.

‡We define $R(C)$ through an integral rather than $R(C) := \mathbb{P}\{C(X) \neq Y\}$ to make it clear that when C is random (depending on training data or random projections), it should be conditioned on when computing $R(C)$.

最小的是贝叶斯估计，也就是在X确定的情况下，Y=1的概率不小于1/2，就归类为1.否则就是0。这个最多就1/2。这么定义R©是因当C随机时，它应该在计算时调节条件。但贝叶斯分类器啥用都没有，因为$\eta(x)$不知道啊。但只用$\eta(x)$的近似值的话，可以通过训练集数据来获取

Bayes classifier. Throughout this section and Section 3, it is convenient to consider the training sample $\mathcal{T}_n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to be fixed points in $\mathbb{R}^p \times \{0, 1\}$. Our methodology will be applied to a base classifier $C_n = C_{n,\mathcal{T}_{n,d}}$ which we assume can be constructed from an arbitrary training sample $\mathcal{T}_{n,d}$ of size n in $\mathbb{R}^d \times \{0, 1\}$; thus C_n is a measurable function from $(\mathbb{R}^d \times \{0, 1\})^n$ to $\mathcal{C}_d$.

这里把分类器的记号重写了下，因为处理的不再是p维的数据，而是n个d维度数据（不含标签）

8. 假设$d \leq p$
9. **投影的定义！**与符号的变更！

Now assume that $d \leq p$. We say a matrix $A \in \mathbb{R}^{d\times p}$ is a *projection* if $AA^T = I_{d\times d}$, the d-dimensional identity matrix. Let $\mathcal{A} = \mathcal{A}_{d\times p} := \{A \in \mathbb{R}^{d\times p} : AA^T = I_{d\times d}\}$ be the set of all such matrices. Given a projection $A \in \mathcal{A}$, define projected data $z_i^A := Ax_i$ and $y_i^A := y_i$ for $i = 1, \ldots, n$, and let $\mathcal{T}_n^A := \{(z_1^A, y_1^A), \ldots, (z_n^A, y_n^A)\}$. The projected data base classifier corresponding to C_n is $C_n^A : (\mathbb{R}^d \times \{0, 1\})^n \to \mathcal{C}_p$, given by

$$C_n^A(x) = C_{n,\mathcal{T}_n^A}^A(x) := C_{n,\mathcal{T}_n^A}(Ax).$$

Note that although C_n^A is a classifier on $\mathbb{R}^p$, the value of $C_n^A(x)$ only depends on x through its d-dimensional projection Ax.

【水字数啊（小声）】就说现在！$C_n^A(x)$虽然是R^p上的分类器，但是人家实际处理的是R^d上哒！

10. 集成要用到的是B_1个投影阵A。在这里投影阵A的分布考虑了训练集数据，然而实际没必要

We now define a generic ensemble classifier based on random projections. For $B_1 \in \mathbb{N}$, let $\mathbf{A}_1, \ldots, \mathbf{A}_{B_1}$ denote independent and identically distributed projections in $\mathcal{A}_{d\times p}$, independent of (X, Y). The distribution on $\mathcal{A}$ is left unspecified at this stage, and in fact our proposed method ultimately involves choosing this distribution depending on $\mathcal{T}_n$.

11.

（1）式是分类器对分类出来结果是1，也分对的情况的频率

（2）式是定义了一个集成分类器，只要（1）式不小于α,，就可以分类为1

总结起来，就是这玩意儿把选出来的是1的标签统计了下，不小于某个频率就可以把这个定成标签1。

$$\nu_n(x) = \nu_n^{(B_1)}(x) := \frac{1}{B_1}\sum_{b_1=1}^{B_1} \mathbb{1}_{\{C_n^{\mathbf{A}_{b_1}}(x)=1\}}. \tag{1}$$

For $\alpha \in (0,1)$, the *random projection ensemble* classifier is defined to be

$$C_n^{\mathrm{RP}}(x) := \begin{cases} 1 & \text{if } \nu_n(x) \geq \alpha; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

这个频率（阈值）不一定是1/2，是数据驱动的软阈值。

11. 这里定义了下期望

$$\mu_n(x) := \mathbf{E}\{\nu_n(x)\} = \mathbf{P}\{C_n^{\mathbf{A}_1}(x) = 1\}.$$

For $r = 0, 1$, define distribution functions $G_{n,r} : [0,1] \to [0,1]$ by $G_{n,r}(t) := P_r(\{x \in \mathbb{R}^p : \mu_n(x) \leq t\})$. Note that since $G_{n,r}$ is non-decreasing it is differentiable almost everywhere; in fact, however, the following assumption will be convenient:

还有这期望可以干嘛用：定义了一个阶梯的概率函数

12. 这概率函数还有啥性质？

- 假设在α处二次可导

Assumption 1. $G_{n,0}$ and $G_{n,1}$ are twice differentiable at α.

under assumption 1, these derivatives are well-defined in a neighbourhood of α. Our first main result below gives an asymptotic expansion for the expected test error $\mathbf{E}\{R(C_n^{\mathrm{RP}})\}$ of our generic random projection ensemble classifier as the number of projections increases. In particular, we show that this expected test error can be well approximated by the test error of the infinite-simulation random projection classifier

$$C_n^{\mathrm{RP}^*}(x) := \begin{cases} 1 & \text{if } \mu_n(x) \geq \alpha; \\ 0 & \text{otherwise.} \end{cases}$$

Note that provided $G_{n,0}$ and $G_{n,1}$ are continuous at α, we have

$$R(C_n^{\mathrm{RP}^*}) = \pi_1 G_{n,1}(\alpha) + \pi_0\{1 - G_{n,0}(\alpha)\}. \tag{3}$$

THEOREM 1. *Assume assumption 1. Then*

$$\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C_n^{\mathrm{RP}^*}) = \frac{\gamma_n(\alpha)}{B_1} + o\Big(\frac{1}{B_1}\Big)$$

as $B_1 \to \infty$, *where*

$$\gamma_n(\alpha) := (1 - \alpha - [\![B_1\alpha]\!])\{\pi_1 g_{n,1}(\alpha) - \pi_0 g_{n,0}(\alpha)\} + \frac{\alpha(1-\alpha)}{2}\{\pi_1 \dot{g}_{n,1}(\alpha) - \pi_0 \dot{g}_{n,0}(\alpha)\}.$$

牛掰。这里在写出某个确定的rpEnsemble分类器后写出了这个分类器的测试误差（3）式。理论一推了下这个分类器的测试误差和全部RPEnsemble分类器的期望的差值，又推出来这个期望与贝叶斯的误差

THEOREM 2. *For each* $B_1 \in \mathbb{N} \cup \{\infty\}$, *we have*

$$\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C^{\mathrm{Bayes}}) \leq \frac{1}{\min(\alpha, 1-\alpha)}\big[\mathbf{E}\{R(C_n^{\mathbf{A}_1})\} - R(C^{\mathrm{Bayes}})\big]. \tag{4}$$

§In order to distinguish between different sources of randomness, we will write $\mathbf{P}$ and $\mathbf{E}$ for the probability and expectation, respectively, taken over the randomness from the projections $\mathbf{A}_1, \ldots, \mathbf{A}_{B_1}$. If the training data is random, then we condition on $\mathcal{T}_n$ when computing $\mathbf{P}$ and $\mathbf{E}$.

当B_1等于无穷时，就把理论2中的替换成带星号的。

When $B_1 = \infty$, we interpret $R(C_n^{\mathrm{RP}})$ in Theorem 2 as $R(C_n^{\mathrm{RP}^*})$. In fact, when $B_1 = \infty$ and $G_{n,0}$ and $G_{n,1}$ are continuous, the bound in Theorem 2 can be improved if one is using an 'oracle' choice of the voting threshold α, namely

$$\alpha^* \in \operatorname*{argmin}_{\alpha' \in [0,1]} R(C_{n,\alpha'}^{\mathrm{RP}^*}) = \operatorname*{argmin}_{\alpha' \in [0,1]} \big[\pi_1 G_{n,1}(\alpha') + \pi_0\{1 - G_{n,0}(\alpha')\}\big], \tag{5}$$

这里找出最优的α，也是通过一个一个去试才知道的。在知道了最优的α以后，

$$R(C_{n,\alpha^*}^{\mathrm{RP}^*}) - R(C^{\mathrm{Bayes}}) \le R(C_{n,1/2}^{\mathrm{RP}^*}) - R(C^{\mathrm{Bayes}}) \le 2\big[\mathbf{E}\{R(C_n^{\mathbf{A}_1})\} - R(C^{\mathrm{Bayes}})\big], \tag{6}$$

可以对其进行提升

which improves the bound in (4) since $2 \le \frac{1}{\min\{\alpha^*,(1-\alpha^*)\}}$. It is also worth mentioning that if assumption 1 holds at $\alpha^* \in (0,1)$, and $G_{n,0}$ and $G_{n,1}$ are continuous, then $\pi_1 g_{n,1}(\alpha^*) = \pi_0 g_{n,0}(\alpha^*)$ and the constant in Theorem 1 simplifies to

$$\gamma_n(\alpha^*) = \frac{\alpha^*(1-\alpha^*)}{2}\{\pi_1 \dot{g}_{n,1}(\alpha^*) - \pi_0 \dot{g}_{n,0}(\alpha^*)\} \ge 0.$$

2.2. 随机投影阵的选择【对应第三节】

2.2.1. 记号

1. R_n^A是对分类器错误的估计的记号
2. 一共有B_1批，每批次有B_2个

in Section 2, where we propose a screening method for choosing the random projections. Let R_n^A be an estimator of $R(C_n^A)$, based on $\{(z_1^A, y_1^A), \ldots, (z_n^A, y_n^A)\}$, that takes values in the set $\{0, 1/n, \ldots, 1\}$. Examples of such estimators include the training error and leave-one-out estimator; we discuss these choices in greater detail in Section 4. For $B_1, B_2 \in \mathbb{N}$, let $\{\mathbf{A}_{b_1,b_2} : b_1 = 1, \ldots, B_1, b_2 = 1, \ldots, B_2\}$ denote independent projections, independent of (X, Y), distributed according to Haar measure on $\mathcal{A}$. One way to simulate from Haar measure on the set $\mathcal{A}$ is to first generate a matrix $\mathbf{Q} \in \mathbb{R}^{d \times p}$, where each entry is drawn independently from a standard normal distribution, and then take $\mathbf{A}^T$ to be the matrix of left singular vectors in the singular value decomposition of $\mathbf{Q}^T$ (see, for example, Chikuse,

2.2.2. 如何生成需要的随机矩阵

1. 生成随机矩阵Q（d*p）。每个元素都是从标准正态分布获取
2. 对Q^T进行SVD分解，A^T就是左奇异向量矩阵

2.2.3. 每批矩阵选取

1.

2003, Theorem 1.5.4). For $b_1 = 1, \ldots, B_1$, let

$$b_2^*(b_1) := \operatorname*{sargmin}_{b_2 \in \{1,\ldots,B_2\}} R_n^{\mathbf{A}_{b_1,b_2}}, \tag{7}$$

where sargmin denotes the smallest index where the minimum is attained in the case of a tie. We now set $\mathbf{A}_{b_1} := \mathbf{A}_{b_1,b_2^*(b_1)}$, and consider the random projection ensemble classifier from Section 2 constructed using the independent projections $\mathbf{A}_1, \ldots, \mathbf{A}_{B_1}$.

就是选出来每批里在测试集（？）上表现最好的B_2。重复B_1次

Let

$$R_n^* := \min_{A \in \mathcal{A}} R_n^A$$

denote the optimal test error estimate over all projections. The minimum is attained here, since R_n^A takes only finitely many values. We assume the following:

测试集上的误差，可以有最小值。因为最小最小是

2. 假设2可以把任何一个随机矩阵分类错误的估计控制在最小误差为中心，距离为某个值的区间上，且落在这个区间上的概率不小于β。

Assumption 2. There exists $\beta \in (0,1]$ such that

$$\mathbf{P}\big(R_n^{\mathbf{A}_{1,1}} \le R_n^* + |\epsilon_n|\big) \ge \beta,$$

where $\epsilon_n = \epsilon_n^{(B_2)} := \mathbf{E}\{R(C_n^{\mathbf{A}_1}) - R_n^{\mathbf{A}_1}\}$.

这个值也很奇妙，是同一批里选出来最佳的那个，本身的误差值和估计的误差值的差的期望，由B_2决定。啊，不过通过B_2的增加（花了更多的时间）来选出最好的投影可能是徒劳的，因为你可能会发现一个误差估计较低的投影，但所选择的投影并不一定会产生一个测试误差较低的分类器。

3. 在这种情况下，以下结果根据基于d维数据的分类器的测试超出风险来控制我们随机投影集成分类器的测试超出风险。某个项反映了我们根据预测数据估计分类器测试误差的能力，和一个依赖于预测的数量的项

THEOREM 3. *Assume assumption 2. Then, for each $B_1, B_2 \in \mathbb{N}$, and every $A \in \mathcal{A}$,*

$$\mathbf{E}\{R(C_n^{\mathrm{RP}})\} - R(C^{\mathrm{Bayes}}) \le \frac{R(C_n^A) - R(C^{\mathrm{Bayes}})}{\min(\alpha, 1-\alpha)} + \frac{2|\epsilon_n| - \epsilon_n^A}{\min(\alpha, 1-\alpha)} + \frac{(1-\beta)^{B_2}}{\min(\alpha, 1-\alpha)}, \qquad (8)$$

where $\epsilon_n^A := R(C_n^A) - R_n^A$.

4.

Assumption 3. There exists a projection $A^* \in \mathcal{A}$ such that

$$P_X(\{x \in \mathbb{R}^p : \eta(x) \ge 1/2\} \triangle \{x \in \mathbb{R}^p : \eta^{A^*}(A^*x) \ge 1/2\}) = 0,$$

where $B \triangle C := (B \cap C^c) \cup (B^c \cap C)$ denotes the symmetric difference of two sets B and C.

PROPOSITION 1. *If Y is conditionally independent of X given A^*X, then assumption 3 holds.*

The following result confirms that under assumption 3, and for a sensible choice of base classifier, we can hope for $R(C_n^{A^*})$ to be close to the Bayes risk.

PROPOSITION 2. *Assume assumption 3. Then $R(C^{A^*-\mathrm{Bayes}}) = R(C^{\mathrm{Bayes}})$.*

We are therefore now in a position to study the first two terms in the bound in Theorem 3 in more detail for specific choices of base classifier.

2.3. 数据集划分【对应第四节】

请注意，LDA，QDA,KNN的说明已跳过。是在理论三中，推导前两个项的期望值上的边界。

in conjunction with a sample splitting strategy. The idea is to split the sample $\mathcal{T}_n$ into $\mathcal{T}_{n,1}$ and $\mathcal{T}_{n,2}$, say, where $|\mathcal{T}_{n,1}| =: n^{(1)}$ and $|\mathcal{T}_{n,2}| =: n^{(2)}$. To estimate the test error of $C_{n^{(1)}}^A$, the projected data base classifier trained on $\mathcal{T}_{n,1}^A := \{(Z_i^A, Y_i^A) : (X_i, Y_i) \in \mathcal{T}_{n,1}\}$, we use

$$R_{n^{(1)},n^{(2)}}^A := \frac{1}{n^{(2)}} \sum_{(X_i,Y_i)\in\mathcal{T}_{n,2}} \mathbb{1}_{\{C_{n^{(1)}}^A(X_i) \ne Y_i\}};$$

也就是使用训练集投影数据产生分类器，用测试集上的错误分类的数据的比例算作测试集的误差

group of projections, we then select a projection $\mathbf{A}_{b_1}$ that minimises this estimate of test error, and construct the random projection ensemble classifier $C_{n^{(1)},n^{(2)}}^{\mathrm{RP}}$ from

$$\nu_{n^{(1)}}(x) := \frac{1}{B_1} \sum_{b_1=1}^{B_1} \mathbb{1}_{\{C_{n^{(1)}}^{\mathbf{A}_{b_1}}(x)=1\}}.$$

Writing $R_{n^{(1)},n^{(2)}}^* := \min_{A\in\mathcal{A}} R_{n^{(1)},n^{(2)}}^A$, we introduce the following assumption analogous to assumption 2:

Assumption 2'. There exists $\beta \in (0,1]$ such that

$$\mathbf{P}\big(R_{n^{(1)},n^{(2)}}^{\mathbf{A}_{1,1}} \le R_{n^{(1)},n^{(2)}}^* + |\epsilon_{n^{(1)},n^{(2)}}|\big) \ge \beta,$$

where $\epsilon_{n^{(1)},n^{(2)}} := \mathbf{E}\{R(C_{n^{(1)}}^{\mathbf{A}_1}) - R_{n^{(1)},n^{(2)}}^{\mathbf{A}_1}\}$.

推出来

COROLLARY 1. *Assume assumptions 2′ and 3. Then, for each* $B_1, B_2 \in \mathbb{N}$,

$$\mathbf{E}\{R(C^{\mathrm{RP}}_{n^{(1)},n^{(2)}})\} - R(C^{\mathrm{Bayes}}) \le \frac{R(C^{A^*}_{n^{(1)}}) - R(C^{A^*-\mathrm{Bayes}})}{\min(\alpha, 1-\alpha)} + \frac{2|\epsilon_{n^{(1)},n^{(2)}}| - \epsilon^{A^*}_{n^{(1)},n^{(2)}}}{\min(\alpha, 1-\alpha)} + \frac{(1-\beta)^{B_2}}{\min(\alpha, 1-\alpha)},$$

where $\epsilon^{A^*}_{n^{(1)},n^{(2)}} := R(C^{A^*}_{n^{(1)}}) - R^{A^*}_{n^{(1)},n^{(2)}}$.

It then follows by very similar arguments to those given in Section 4.1 that

$$\mathbb{E}(|\epsilon^{A^*}_{n^{(1)},n^{(2)}}| \,|\, \mathcal{T}_{n,1}) = \mathbb{E}\{|R(C^{A^*}_{n^{(1)}}) - R^{A^*}_{n^{(1)},n^{(2)}}| \,|\, \mathcal{T}_{n,1}\} \le \Big(\frac{1+\log 2}{2n^{(2)}}\Big)^{1/2},$$
$$\mathbb{E}(|\epsilon_{n^{(1)},n^{(2)}}| \,|\, \mathcal{T}_{n,1}) = \mathbb{E}\{|R(C^{\mathbf{A}_1}_{n^{(1)}}) - R^{\mathbf{A}_1}_{n^{(1)},n^{(2)}}| \,|\, \mathcal{T}_{n,1}\} \le \Big(\frac{1+\log 2+\log B_2}{2n^{(2)}}\Big)^{1/2}. \quad (16)$$

2.4. 参数选择【对应第五节】

2.4.1 阈值 α

π_0 and π_1 either). Nevertheless, for the LDA base classifier we can estimate $G_{n,r}$ using

$$\hat{G}_{n,r}(t) := \frac{1}{n_r} \sum_{\{i:Y_i=r\}} \mathbb{1}_{\{\nu_n(X_i)<t\}}$$

for $r = 0, 1$. For the QDA and k-nearest neighbour base classifiers, we use the leave-one-out-based estimate $\tilde{\nu}_n(X_i) := B_1^{-1} \sum_{b_1=1}^{B_1} \mathbb{1}_{\{C^{\mathbf{A}_{b_1}}_{n,i}(X_i)=1\}}$ in place of $\nu_n(X_i)$. We also estimate π_r by $\hat{\pi}_r := n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=r\}}$, and then set the cut-off in (2) as

$$\hat{\alpha} \in \operatorname*{argmin}_{\alpha' \in [0,1]} \big[\hat{\pi}_1 \hat{G}_{n,1}(\alpha') + \hat{\pi}_0\{1 - \hat{G}_{n,0}(\alpha')\}\big]. \quad (17)$$

因为（17）式没有确定的最小值（这些都是从实验中得到的），所以把α的估计值取成最小和最大的最小值的均值。不用花多少时间，因为已经算过了

2.4.2. B_1和B_2的选择

推荐$B_1 = 500$，$B_2 = 50$

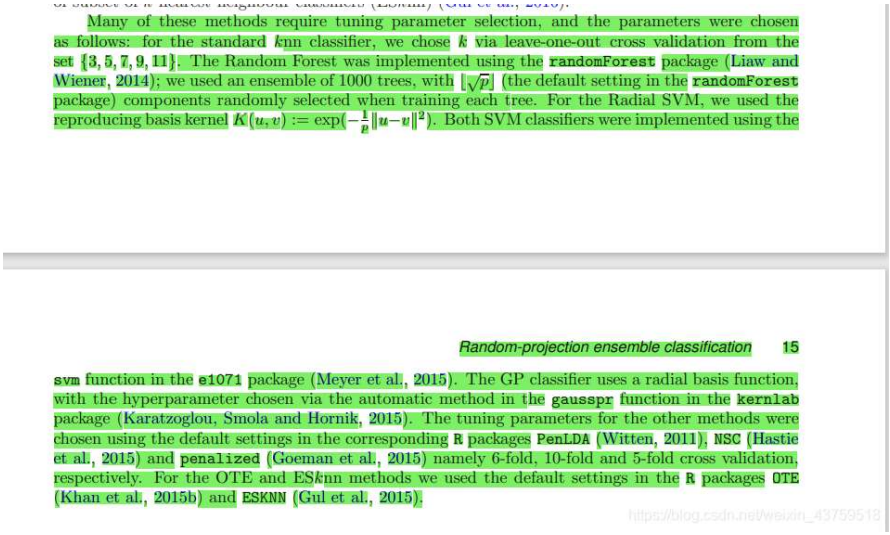
2.4.3. d的选择

constrained. Thus, if we wish to choose d from a set $\mathcal{D} \subseteq \{1, \ldots, p\}$, then for each $d \in \mathcal{D}$, we train the random projection ensemble classifier, and set

$$\hat{d} := \operatorname*{sargmin}_{d \in \mathcal{D}} \big[\hat{\pi}_1 \hat{G}_{n,1}(\hat{\alpha}) + \hat{\pi}_0\{1 - \hat{G}_{n,0}(\hat{\alpha})\}\big],$$

where $\hat{\alpha} = \hat{\alpha}_d$ is given in (17). Such a proceedure does not add to the computational cost at *test time*.

2.5. 实验【对应第六节】



Many of these methods require tuning parameter selection, and the parameters were chosen as follows: for the standard $k$nn classifier, we chose k via leave-one-out cross validation from the set $\{3, 5, 7, 9, 11\}$. The Random Forest was implemented using the `randomForest` package (Liaw and Wiener, 2014); we used an ensemble of 1000 trees, with $\lfloor\sqrt{p}\rfloor$ (the default setting in the `randomForest` package) components randomly selected when training each tree. For the Radial SVM, we used the reproducing basis kernel $K(u, v) := \exp(-\frac{1}{p}\|u-v\|^2)$. Both SVM classifiers were implemented using the

Random-projection ensemble classification 15

`svm` function in the `e1071` package (Meyer et al., 2015). The GP classifier uses a radial basis function, with the hyperparameter chosen via the automatic method in the `gausspr` function in the `kernlab` package (Karatzoglou, Smola and Hornik, 2015). The tuning parameters for the other methods were chosen using the default settings in the corresponding `R` packages `PenLDA` (Witten, 2011), `NSC` (Hastie et al., 2015) and `penalized` (Goeman et al., 2015) namely 6-fold, 10-fold and 5-fold cross validation, respectively. For the OTE and ES$k$nn methods we used the default settings in the `R` packages `OTE` (Khan et al., 2015b) and `ESKNN` (Gul et al., 2015).

其他模型的参数设置

2.5.1. 模拟数值实验

简单来说，就是使用

1. $n = 50, 200, 1000$个数据
2. 每个数据可有$p = 100, 1000$维
3. 两种不同的先验概率
4. 使用高斯投影，一共500批，每批50个
   表格1&2的投影就是使用维度为100的，$Y = 1$占了一半的。其中：

- 测试集大小为1000
- 实验重复100次

2.5.1.1. 稀疏类边界【模型1】

d = 2.作者说该实验聊胜于无……

2.5.1.2. 旋转稀疏正交【模型2】

d =3.和模型1类似，不过进行了旋转。也不是单纯的对角矩阵，非对角线元素也不一定是0.RPEnsemble在这里，不论d是过小还是过大，表现都不错

2.5.1.3. 独立特征【模型3】

这个模型的类边界是非线性的。假设3不一定适合所有d<p的模型。作者还说：
and in fact the RP-QDA5 classifier has the smallest misclassification rate among all methods implemented
因为非线性边界的自然不适合用LDA搞

2.5.1.4. t分布特征【模型4】

模型4探讨了重尾的影响以及特征之间存在的相关性。套上LDA和QDA效果也不好，可能是因为：
类条件分布的二阶矩和一阶矩分别是有限的，因此，类均值和协方差矩阵估计很差

2.5.2. 真实数据集

这些数据集来自UCI
使用n个作为训练集，min(剩下的,1000)作为测试集。一共500批，每批50个。使用高斯投影。实验重复100次。实验是经过调参的【2333】

- Eye state detection【p=14】
- Ionosphere dataset【p=32】

- Down's syndrome diagnoses in mice【p=77】
- Hill-Valley identification【p=100】
- Musk identification【p=166】
- Cardiac Arrhythmia diagnoses【p= 194】
- Human Activity Recognition【p=561】
- Handwritten digits/Gisette dataset【p=5000】
  第6.3节总结了下，有多少个数据集里，本算法是最好的；或者排名前三的；还吹了吹普适性；
  当然，假设三不是必须品，已经强调好几次了。
  在Gisette dataset上这个框架不够有效。因为太稀疏了。用随机投影可能会破坏稀疏结构
  【小声，李平的CRP应该可以解决这个问题】这个时候可以使用投影的替代分布，例如轴向对齐的投影

3. 以上都是二分类数据集

4.有一点没想明白，如何知道是测试集上最好的?

通过在训练集/验证集上的表现选出，并祈祷它们是最好