

基于类别随机化的随机森林算法

关晓蔷 庞继芳 梁吉业

(山西大学计算机与信息技术学院 太原 030006)

(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)

摘 要 随机森林是数据挖掘和机器学习领域中一种常用的分类方法,已成为国内外学者共同关注的研究热点,并被广泛应用到各种实际问题中。传统的随机森林方法没有考虑类别个数对分类效果的影响,忽略了基分类器和类别之间的关联性,导致随机森林在处理多分类问题时的性能受到限制。为了更好地解决该问题,结合多分类问题的特点,提出一种基于类别随机化的随机森林算法(RCRF)。从类别的角度出发,在随机森林两种传统随机化的基础上增加类别随机化,为不同类别设计具有不同侧重点的基分类器。由于不同的分类器侧重区分的类别不同,所生成的决策树的结构也不同,这样既能够保证单个基分类器的性能,又可以进一步增大基分类器的多样性。为了验证所提算法的有效性,在 UCI 数据库中的 21 个数据集上将 RCRF 与其他算法进行了比较分析。实验从两个方面进行,一方面,通过准确率、F1-measure 和 Kappa 系数 3 个指标来验证 RCRF 算法的性能;另一方面,利用 κ -误差图从多样性角度对各种算法进行对比与分析。实验结果表明,所提算法能够有效提升集成模型的整体性能,在处理多分类问题时具有明显优势。

关键词 随机森林,多分类问题,类别随机化,多样性

中图法分类号 TP181 文献标识码 A DOI 10.11896/j.issn.1002-137X.2019.02.030

Randomization of Classes Based Random Forest Algorithm

GUAN Xiao-qiang PANG Ji-fang LIANG Ji-ye

(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

(Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University),
Ministry of Education, Taiyuan 030006, China)

Abstract Random forest is a commonly used classification method in the field of data mining and machine learning, which has become a research focus of scholars at home and abroad, and has been widely applied to various practical problems. The traditional random forest methods do not consider the influence of the number of classes on the classification effect, and neglect the correlation between base classifiers and classes, limiting the performance of the random forest in dealing with multi-class classification problems. In order to solve the problem better, combined with the characteristics of multi-class classification problem, this paper proposed a randomization of classes based random forest algorithm (RCRF). From the perspective of classes, the randomization of classes is added on the basis of two kinds of traditional randomizations of random forest, and the corresponding base classifiers with different emphasis are designed for different classes. The structures of the decision tree generated by the base classifier are different because different classifiers focus on different classes, which can not only guarantee the performance of the single base classifier, but also further increase the diversity of base classifier. In order to verify the validity of the proposed algorithm, RCRF is compared with other algorithms on 21 data sets in UCI database. The experiment is carried out from two aspects. On the one hand, the accuracy, F1-measure and Kappa coefficient are used to verify the performance of RCRF algorithm. On the other hand, the κ -error diagram is used to compare and analyze various algorithms from the perspective of diversity. Experimental results show that the proposed algorithm can effectively improve the overall performance of the integrated model and has obvious advantages in dealing with multi-class classification problems.

Keywords Random forest, Multi-class classification problems, Randomization of classes, Diversity

到稿日期:2018-09-07 返修日期:2018-11-23 本文受国家自然科学基金项目(61876103),山西省青年科技基金项目(201701D221098),山西省重点研发项目(201603D111014),山西省留学基金项目(2016-003)资助。

关晓蔷(1979-),女,博士生,讲师,CCF会员,主要研究方向为数据挖掘与机器学习,E-mail:gxq0079@sxu.edu.cn;庞继芳(1980-),女,博士,讲师,CCF会员,主要研究方向为智能决策与数据挖掘;梁吉业(1962-),男,博士,教授,CCF会员,主要研究方向为粒计算、数据挖掘与机器学习,E-mail:lji@sxu.edu.cn(通信作者)。

1 引言

分类是数据挖掘领域中一项非常重要的任务,解决分类问题的方法(即分类算法)是机器学习和模式识别中一个重要的研究方向。典型的分类算法包括随机森林(Random Forests, RF)^[1]、神经网络、支持向量机、贝叶斯网络等。随机森林通过集成学习的思想对多棵决策树的结果进行集成,是一种基于数据驱动的非参数分类方法。大量的理论和实验研究表明,随机森林具有很高的预测准确率,不容易出现过拟合现象,其性能往往优于其他学习方法^[2]。因此,随机森林得到了学者们的青睐,已成为数据分析和挖掘、知识管理、模式识别等众多领域的研究热点,并被广泛应用到如生物信息^[3]、医学研究^[4-5]、人脸图像识别^[6]、目标跟踪^[7]等实际问题中。

近年来,针对随机森林的相关研究主要集中在算法性能的提升方面。Geurts 等提出了一种极端随机树算法^[8],该算法摒弃了传统的 bootstrap 采样方法,直接使用原始的训练本来减小偏差,并在每棵决策树的决策节点上随机选择分裂测试的阈值,使得极端随机森林分类器在分类精度和训练时间方面都优于随机森林分类器。旋转森林(Rotation Forest, ROF)^[9]是 Rodriguez 等基于特征变换思想提出的一种集成算法,该算法专注于提高基分类器的差异性和准确性。Zhang 等在旋转森林的基础上提出基于随机特征空间的随机森林方法^[10],该方法通过在每个节点上应用 PCA 或 LDA 进行特征变换来进一步提高随机森林的性能。Abellán 等^[11]利用不精确信息增益作为属性选择的标准,建立了基于不精确概率理论的 Credal Random Forest(CRF),实验证明该算法在提高随机森林性能的同时,对有类别噪音的数据进行分类时的效果也较明显。Wang 等^[12]为了缩小随机森林理论在一致性和实际性能之间的差距,建立了伯努利随机森林(BRFs)。Ye 等^[13]为了保证高维数据下每个随机子空间都包含足够多的有用信息,设计了一种分层抽样的方法来选择随机子空间。按照分类时提供信息量的多少将特征分成两个子集,然后从两个子集中随机抽取特征,从而有效地解决高维数据下的分类问题。针对不完整数据的分类任务,Xia 等^[14]通过估计缺失数据对决策树的影响来调整每棵决策树的投票权重,从而给出了一种加权投票随机森林(AWVRF)算法。Hu 等^[15]基于分离轴定理(Separating Axis Theorem, SAT)的分割策略,设计了一种增量随机森林(CIRF),该方法可以在不重建子树的情况下插入新结点,实验结果表明,在大多数情况下该方法的测试准确率都优于其他增量学习算法。

上述研究成果丰富了随机森林的相关理论,为随机森林的进一步发展提供了新的思路。现有的随机森林方法没有考虑类别个数对分类效果的影响,忽略了基分类器和类别之间的关联性,导致随机森林在处理多分类问题时的性能受到限制。在多分类问题中,一个样本属于且只属于多个类中的一个,且不同的类之间是互斥的。多分类方法旨在通过对不同类的样本进行训练来实现对各种未知样本的类的识别。本文

将在已有研究的基础上,结合多分类问题的特点,通过引入类别随机化给出一种适用于多分类问题的随机森林算法。该算法在考虑类别随机化的基础上,针对不同类别训练相应的基分类器,增大基分类器的多样性,从而达到提高随机森林整体性能的目的。

2 随机森林相关理论

随机森林是由 Breiman^[1]于 2001 年首次提出的一种高度灵活的机器学习算法,该算法以统计学习理论为基础,将 bagging^[16]集成学习理论与随机子空间方法^[17]相结合,利用 bootstrap 重抽样方法从原始样本中抽取多个样本,在对每个 bootstrap 样本进行决策树建模的基础上,组合多棵决策树的预测,进而通过投票得出最终的预测结果。

众所周知,随机森林的性能主要取决于两方面的因素:1)单个分类器的精度;2)基分类器的多样性。随机森林大都通过训练数据集随机化和属性集随机化两种方式增强树的多样性。首先,对初始训练集数据进行随机放回抽样来生成多个新的训练数据集,新的训练数据集的大小与初始数据集的大小相同。在生成新的训练数据集之后,通过对所有属性进行随机抽样来创建随机属性集以进一步增强其树的多样性,并从随机属性集中选择最佳分割属性对树中的每个结点进行划分。由于训练数据集和属性集都是随机生成的,因此随机森林中树的生长是独立的且互不相同。随机森林通过平均各个决策树的预测来得到最终的预测结果,这种联合预测方法可以有效降低模型的泛化误差。随机森林中每一棵决策树的生成过程如下:

步骤 1 训练数据抽样。设原始样本集的大小为 n ,从原始样本集中随机放回地抽取 n 个样本作为新的训练集。

步骤 2 属性子空间抽样。随机地从 M 个原始属性中选取 m 个属性形成新的属性子空间。

步骤 3 决策树模型的建立。根据 CART 算法构建树,所有树都自然生长,不进行剪枝。

当使用初始训练集的 bootstrap 抽样作为训练集构造决策树时,有一些样本是不会被抽取的,这些样本的个数占初始数据集的 $(1-1/n)^n$ 。可以证明,当 n 足够大时, $(1-1/n)^n$ 将收敛于 $1/e \approx 0.368$,这个数据表明,有将近 37% 的样本未被抽出来,称由这些样本组成的集合为袋外数据,简记为 OOB 数据。在随机森林中,OOB 数据常被用来估计算法的泛化能力。OOB 估计是高效的,其结果近似于需要大量计算的 k 折交叉验证。

3 基于类别随机化的随机森林算法

在保证单棵决策树精度的基础上,创建一组多样化的决策树有助于实现模型互补,提高集成模型的整体预测性能。传统的随机森林大都是通过随机化训练数据集和属性集来增强树的多样性。本文从类别的角度出发,在两种已有随机化的基础上增加了类别的随机化,提出了一种新的随机森林算

法(RCRF)。该算法针对不同类别设计相应的基分类器,在建立基分类器的过程中,随机地从所有类别中选取一个作为优先训练类别。由于不同的分类器侧重的类别不同,针对不同类生成的决策树的结构也大不相同,这样就进一步增加了基分类器的多样性。

给定一个训练集 (X, Y) , $X = \{x_1, x_2, \dots, x_n\}$ 是样本集合,其中 $x_j \in R^M$ 是样本集中的第 j 个样本, $\{\omega_1, \omega_2, \dots, \omega_c\}$ 是样本的类别标签集合, $Y = \{y_1, y_2, \dots, y_n\}$ 是与 X 相对应的类别标签向量,其中 $y_j \in \{\omega_1, \omega_2, \dots, \omega_c\}$ 是样本 x_j 的类别标签。用 RCRF 算法生成分类器时,对于森林中的第 i 棵决策树 T_i ,首先从大小为 n 的训练集 (X, Y) 中随机可放回地抽取 n 个样本作为新的训练集 (X_i', Y_i') ,并从 M 个原始属性中随机地选取 $m (m \ll M)$ 个属性形成新的属性子空间,然后从所有类别 $\{\omega_1, \omega_2, \dots, \omega_c\}$ 中随机抽取一个类别标签 $\omega_k (1 \leq k \leq c)$ 作为决策树 T_i 的优先训练类别。针对类别 ω_k ,为训练集 (X_i', Y_i') 建立一个二值类别标签向量:

$$Y_i^k = \{y_{i1}^k, y_{i2}^k, \dots, y_{in}^k\}$$

$$\text{其中, } y_{ij}^k = \begin{cases} 1, & y'_{ij} = \omega_k \\ 0, & y'_{ij} \neq \omega_k \end{cases}$$

决策树 T_i 中每个节点的分裂分为两种情况。当结点中包含属于类别 ω_k 的样本时,将二值类别标签 Y_i^k 作为该结点样本集的类别标签,进而从 m 个候选属性中选择最佳分裂属性生成子结点;当结点中不包含属于类别 ω_k 的样本时,仍使用原有的标签向量 Y_i^k ,并从 m 个候选属性中选择最佳分裂属性生成子结点。

在测试阶段,分别用 RCRF 模型中生成的 L 棵决策树对测试样本 x 进行预测,每棵决策树 T_i 会给出一个预测结果 $T_i(x)$,从而得到 L 个预测结果 $T_1(x), T_2(x), \dots, T_L(x)$,测试样本 x 最终的类别标签 y 通过式(1)以投票方式给出,得票最多的类别即为随机森林的输出结果。

$$y = \arg \max_{\omega_j} \sum_{i=1}^L I(T_i(x) = \omega_j), j = 1, 2, \dots, c \quad (1)$$

其中, $I(\cdot)$ 是指示函数,当 $T_i(x) = \omega_j$ 时, $I(T_i(x) = \omega_j)$ 的值为 1,否则其值为 0。

RCRF 的算法框架如图 1 所示。算法 1 描述了 RCRF 的整体实现过程,其中单棵决策树的生成过程如算法 2 所示。

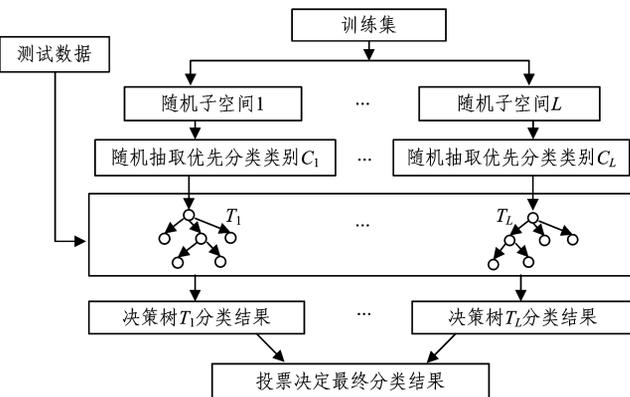


图 1 RCRF 算法框架图

Fig. 1 Framework of RCRF algorithm

算法 1 RCRF 算法

输入:训练集,随机森林中决策树的棵数 L ,测试样本 x

输出:测试样本 x 的类别标签 $y (y \in \{\omega_1, \omega_2, \dots, \omega_c\})$

步骤 1 for $i = 1, 2, \dots, L$

从大小为 n 的训练集 (X, Y) 中随机可放回抽取 n 个样本作为新的训练集 (X_i', Y_i') ;

从 $\{\omega_1, \omega_2, \dots, \omega_c\}$ 中随机抽取一个类别标签 $\omega_k (1 \leq k \leq c)$ 作为第 i 个分类器的优先训练类别;

针对类别 ω_k ,为训练集 (X_i', Y_i') 建立一个二值类别标签

向量 $Y_i^k = \{y_{i1}^k, y_{i2}^k, \dots, y_{in}^k\}$, 其中 $y_{ij}^k = \begin{cases} 1, & y'_{ij} = \omega_k \\ 0, & y'_{ij} \neq \omega_k \end{cases}$;

调用算法 2 建立一棵决策树 T_i 。

end for;

步骤 2 分别用 L 棵决策树对测试样本 x 进行预测,得到 L 个预测结果 $T_1(x), T_2(x), \dots, T_L(x)$,进而通过式(1)以投票方式得到测试样本 x 的类别标签 y 。

算法 2 RCRF 中单棵决策树的生成算法

输入:训练集 (X_i', Y_i^k) , 优先训练类别 ω_k , 候选属性集大小 m

输出:决策树 T_i

步骤 1 创建一个新的结点,包含所有训练样本;

步骤 2 if 结点满足停止分裂条件

将该结点标记为叶子结点,将结点中样本数最多的类作为该结点的类别标签,转步骤 5;

endif

步骤 3 从 M 个原始属性中随机地选取 $m (m \ll M)$ 个属性;

步骤 4 if 结点中包含属于类别 ω_k 的样本

使用算法 1 中建立的二值类别标签 Y_i^k ,从 m 个候选属性中选择最佳分裂属性生成子结点;

else

使用原有的类别标签 Y_i^k ,从 m 个候选属性中选择最佳分裂属性生成子结点;

endif

步骤 5 重复执行步骤 2—步骤 4,直至所有结点都被训练过或被标记为叶子结点。

在算法 2 的步骤 2 中提到的结点的停止分裂条件包含以下两种:1)当前结点包含的样本都属于同一类别;2)当前属性集为空,或所有样本在所有属性上取值相同。步骤 4 中的最佳分裂属性的选择标准可以是信息增益、信息增益率或 Gini 系数等已有准则。

4 实验分析

为了验证所提算法的有效性,本文在 UCI 数据库中的 21 个数据集上将其与已有的随机森林算法 RF^[1] 和 CRF^[11] 进行了比较分析。实验从两个方面进行:一方面,通过准确率、F1-measure 和 Kappa 系数 3 个指标来验证 RCRF 算法的性能;另一方面,利用 κ -误差图从多样性角度对 3 种算法进行对比分析。实验中所有的算法均使用 Matlab2013 实现。

4.1 数据集

表 1 简要描述了本文所选取的 UCI 数据库中的 21 个数据集的大致情况,包括样本数、属性数和类别数。

表 1 UCI 数据集
Table 1 UCI Data set

数据集	样本数	属性数	类别数
balance-scale	625	4	3
car	1728	6	4
ecoli	336	7	8
glass	214	9	7
krkopt	28056	6	18
letter	20000	16	26
mfeat-fac	2000	216	10
mfeat-fou	2000	76	10
mfeat-kar	2000	64	10
mfeat-mor	2000	6	10
mfeat-pix	2000	240	10
mfeat-zer	2000	47	10
nursery	12960	8	5
pendigits	10992	16	10
segment	2310	19	7
shuttle	58000	9	7
tae	151	5	3
texture	5500	40	11
tumor	132	17	18
wine	178	13	3
zoo	101	16	7

4.2 度量指标

本文采用准确率、F1-measure 和 Kappa 系数 3 个指标来评价分类器的性能。

(1) 准确率

准确率(Accuracy)是分类任务中最常用的性能度量指标,表示分类正确的样本数占样本总数的比例,计算式如下:

$$acc = \frac{1}{n} \sum_{i=1}^n I(f(x_i) = y_i) \quad (2)$$

(2) F1-measure

准确率虽然常用,但并不能满足所有的任务需求,尤其对于一些不平衡的样本集来说,一些小类样本虽然数量很少但是却很重要,对它们的错误估计往往会产生较大的实际影响,难以通过准确率反映出来,这时用查准率和查全率能做出更有效的评价。F1-measure 度量是查准率和查全率具有相同权重时的加权调和平均,计算式如下:

$$F1 = \frac{2PR}{P+R} \quad (3)$$

其中, P 是查准率, R 是查全率。

(3) Kappa 系数

Kappa 系数用于一致性检验,可以评估分类模型结果与实际结果的一致性程度,具体公式如下:

$$\kappa = \frac{P_1 - P_2}{1 - P_2} \quad (4)$$

其中, P_1 是两个分类器取得一致的概率, P_2 是两个分类器偶然达成一致的的概率。

$$P_1 = \sum_k m_{k,k} \quad (5)$$

$$P_2 = \sum_k (\sum_s m_{k,s}) (\sum_s m_{s,k}) \quad (6)$$

其中, $m_{k,s}$ 表示分类器 T_i 把测试样本分在 ω_k 类且分类器 T_j 把样本分在 ω_s 类的概率。Kappa 系数的取值范围为 $[-1, 1]$, 其值越大,表明一致性越好。

4.3 性能分析

本节通过对 RCRF 与已有的 RF、CRF 算法在准确率、

F1-measure、Kappa 系数 3 种评价指标上的分析和比较来验证所提算法的性能,其中 RCRF 使用 Gini 系数作为划分属性选择的标准。在实验过程中,集成规模 $L=100$, 候选属性集大小 $m = \log_2(M)$, 并使用 OOB 数据来估计算法的泛化能力。为了保证算法的稳定性,在每一个数据集上将算法重复执行 10 次。3 种算法在 3 个评价指标上的实验结果分别如表 2—表 4 所列。从 3 个表中的数据可以看出,RCRF 在准确率、F1-measure、Kappa 系数 3 个指标上的效果明显优于其他两种算法。

表 2 3 种算法在数据集上的准确率

Table 2 Accuracy of three algorithms on data sets

数据集	RF	CRF	RCRF
balance-scale	0.8310±0.0043	0.8318±0.0069	0.8456±0.0078
car	0.9828±0.0026	0.9815±0.0020	0.9822±0.0021
ecoli	0.8586±0.0098	0.8548±0.0071	0.8664±0.0099
glass	0.7822±0.0125	0.7879±0.0133	0.7921±0.0147
krkopt	0.8079±0.0016	0.7965±0.0014	0.8205±0.0023
letter	0.9641±0.0011	0.9639±0.0007	0.9677±0.0011
mfeat-fac	0.9663±0.0017	0.9652±0.0021	0.9671±0.0027
mfeat-fou	0.8246±0.0053	0.8259±0.0031	0.8249±0.0046
mfeat-kar	0.9528±0.0033	0.9517±0.0026	0.9525±0.0036
mfeat-mor	0.6991±0.0046	0.6997±0.0039	0.7000±0.0048
mfeat-pix	0.9751±0.0023	0.9738±0.0021	0.9742±0.0018
mfeat-zer	0.7650±0.0037	0.7660±0.0041	0.7678±0.0054
nursery	0.9923±0.0009	0.9883±0.0008	0.9926±0.0006
pendigits	0.9915±0.0003	0.9913±0.0004	0.9918±0.0004
segment	0.9799±0.0018	0.9812±0.0015	0.9802±0.0010
shuttle	0.9999±0.0000	0.9999±0.0000	0.9999±0.0000
tae	0.6291±0.0212	0.6444±0.0193	0.6450±0.0219
texture	0.9782±0.0009	0.9792±0.0011	0.9793±0.0009
tumor	0.4212±0.0172	0.4189±0.0208	0.4235±0.0209
wine	0.9781±0.0056	0.9787±0.0036	0.9820±0.0058
zoo	0.9525±0.0112	0.9564±0.0083	0.9584±0.0078
Avg. Rank	2.33	2.38	1.29
Friedman $\alpha=0.05$	✓	✓	—

表 3 3 种算法在数据集上的 F1-measure

Table 3 F1-measure of three algorithms on data sets

数据集	RF	CRF	RCRF
balance-scale	0.5991±0.0022	0.5970±0.0033	0.6030±0.0036
car	0.9486±0.0080	0.9455±0.0067	0.9446±0.0085
ecoli	0.6089±0.0162	0.5963±0.0128	0.6124±0.0135
glass	0.7515±0.0146	0.7530±0.0203	0.7545±0.0258
krkopt	0.8109±0.0026	0.8049±0.0040	0.8206±0.0029
letter	0.9640±0.0011	0.9639±0.0007	0.9676±0.0011
mfeat-fac	0.9663±0.0017	0.9652±0.0021	0.9670±0.0027
mfeat-fou	0.8232±0.0055	0.8243±0.0034	0.8237±0.0046
mfeat-kar	0.9527±0.0033	0.9517±0.0026	0.9524±0.0036
mfeat-mor	0.6971±0.0048	0.6972±0.0040	0.7000±0.0048
mfeat-pix	0.9751±0.0022	0.9738±0.0020	0.9742±0.0018
mfeat-zer	0.7623±0.0036	0.7635±0.0039	0.7652±0.0052
nursery	0.9922±0.0012	0.9883±0.0018	0.9914±0.0013
pendigits	0.9916±0.0003	0.9914±0.0004	0.9919±0.0004
segment	0.9799±0.0018	0.9812±0.0015	0.9802±0.0010
shuttle	0.9675±0.0059	0.9562±0.0080	0.9707±0.0095
tae	0.6255±0.0219	0.6413±0.0203	0.6429±0.0231
texture	0.9782±0.0009	0.9792±0.0011	0.9793±0.0009
tumor	0.2126±0.0115	0.2121±0.0207	0.2158±0.0192
wine	0.9785±0.0054	0.9791±0.0036	0.9822±0.0056
zoo	0.8871±0.0324	0.8943±0.0221	0.9047±0.0190
Avg. Rank	2.29	2.38	1.33
Friedman $\alpha=0.05$	✓	✓	—

表4 3种算法在数据集上的 Kappa 系数

Table 4 Kappa of three algorithms on data sets

数据集	RF	CRF	RCRF
balance-scale	0.7019±0.0070	0.7016±0.0111	0.7238±0.0126
car	0.9625±0.0057	0.9597±0.0043	0.9612±0.0047
ecoli	0.8036±0.0137	0.7981±0.0103	0.8141±0.0138
glass	0.6995±0.0172	0.7071±0.0185	0.7115±0.0210
krkopt	0.7851±0.0018	0.7725±0.0015	0.7993±0.0026
letter	0.9627±0.0012	0.9625±0.0007	0.9664±0.0012
mfeat-fac	0.9626±0.0019	0.9613±0.0024	0.9634±0.0030
mfeat-fou	0.8051±0.0059	0.8066±0.0035	0.8054±0.0051
mfeat-kar	0.9475±0.0036	0.9463±0.0029	0.9472±0.0040
mfeat-mor	0.6657±0.0051	0.6664±0.0044	0.6667±0.0054
mfeat-pix	0.9723±0.0025	0.9713±0.0022	0.9713±0.0020
mfeat-zer	0.7388±0.0041	0.7401±0.0046	0.7419±0.0060
nursery	0.9887±0.0013	0.9828±0.0011	0.9891±0.0009
pendigits	0.9906±0.0003	0.9904±0.0004	0.9909±0.0004
segment	0.9766±0.0021	0.9780±0.0018	0.9769±0.0012
shuttle	0.9996±0.0000	0.9996±0.0000	0.9996±0.0001
tae	0.4441±0.0317	0.4670±0.0289	0.4678±0.0328
texture	0.9760±0.0010	0.9771±0.0012	0.9772±0.0010
tumor	0.3426±0.0190	0.3390±0.0243	0.3431±0.0247
wine	0.9668±0.0085	0.9676±0.0054	0.9728±0.0088
zoo	0.9372±0.0148	0.9424±0.0110	0.9451±0.0103
Avg. Rank	2.29	2.40	1.31
Friedman $\alpha=0.05$	✓	✓	—

为了进一步验证 RCRF 算法是否在统计学上显著优于其他方法,本文在 21 个数据集上使用基于算法排序的 Friedman 检验^[18]对上述算法进行比较。在每个数据集上,针对上述算法,根据测试性能由好到坏排序,并赋予序值 1,2,⋯,若算法的测试性能相同,则平分序值。最后通过对每一列的序值求平均,来得到平均序值,并在表格的底部标出了显著性分析结果。“✓”表明在显著性水平 $\alpha=0.05$ 下 RCRF 算法在统计学上显著优于该算法。

由表 2—表 4 可以看出,当 $\alpha=0.05$ 时,RCRF 算法在准确率、F1 值、Kappa 系数 3 种评价指标的性能统计上均显著优于 RF 和 CRF 算法。

4.4 κ -误差图

为了进一步分析所提算法的多样性,本文还使用了 Margineantu 等提出的 κ -误差图^[19]。 κ -误差图主要用于测量两个分类器输出的成对多样性以及它们的平均误差,其中成对多样性通过 Kappa(κ)来度量,采用 10 次 10 折交叉验证。当集成规模为 L 时,将产生 $L(L-1)/2$ 对分类器,每一对分类器构成图中的一个点,每个点的横坐标是这对分类器的 Kappa 值,纵坐标是这对分类器的平均误差。由于空间限制,本文仅画出了 3 种算法在 krkopt 数据集上的 κ -误差图,如图 2 所示。3 种集成分类器中决策树的数目 L 均为 100,因此图 2 中每个子图都有 4950 个点。数据点云的位置越靠上,个体分类器的准确性就越低;数据点云的位置越靠右, κ 值越大,成对分类器的一致性程度越高,表明个体分类器的多样性越小。图 3 给出了 3 种算法在 krkopt 数据集上的 κ -误差图的中心点, x 轴是成对分类器 kappa 值的均值, y 轴是成对分类器平均误差的均值。通过图 3 可以更加直观地对各种集成算法的数据点云的相对位置进行比较和评估。所有数据集的 κ -误差的均值数据如表 5 所列。

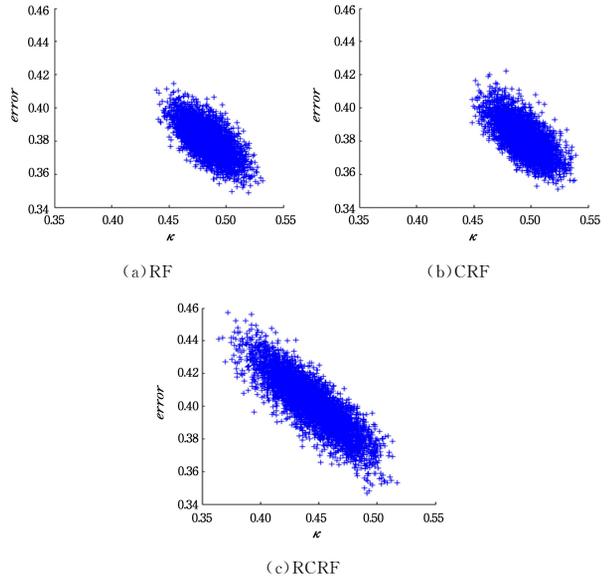


图2 数据集 krkopt 的 κ -误差图

Fig. 2 κ -error diagrams of krkopt data set

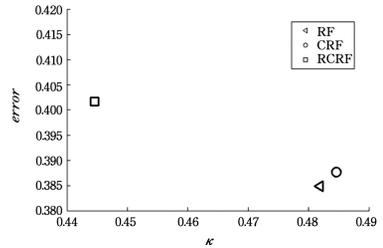


图3 数据集 krkopt 的 κ -误差图中心

Fig. 3 Centroids of κ -error diagrams of krkopt data set

表5 3种算法的 κ -误差数据点

Table 5 κ -error data of three algorithms

数据集	RF		CRF		RCRF	
	κ	error	κ	error	κ	error
balance-scale	0.5721	0.2375	0.5657	0.2369	0.5489	0.2404
car	0.7266	0.0856	0.6948	0.0944	0.6550	0.1060
ecoli	0.6532	0.2362	0.6548	0.2382	0.6264	0.2454
glass	0.4347	0.3641	0.4423	0.3573	0.4015	0.3760
krkopt	0.4819	0.3849	0.4846	0.3877	0.4446	0.4017
letter	0.7344	0.1781	0.7421	0.1733	0.7022	0.1970
mfeat-fac	0.6947	0.1835	0.7013	0.1798	0.6552	0.2055
mfeat-fou	0.4377	0.3955	0.4488	0.3880	0.3685	0.4420
mfeat-kar	0.4296	0.3529	0.4442	0.3430	0.3571	0.4029
mfeat-mor	0.6830	0.3383	0.6815	0.3356	0.6713	0.3428
mfeat-pix	0.6601	0.1988	0.6763	0.1891	0.6109	0.2264
mfeat-zer	0.5107	0.3931	0.5217	0.3880	0.4510	0.4281
nursery	0.8325	0.0697	0.7720	0.0961	0.8046	0.0808
pendigits	0.8935	0.0605	0.9012	0.0563	0.8786	0.0678
segment	0.8916	0.0652	0.8959	0.0626	0.8740	0.0739
shuttle	0.9973	0.0007	0.9975	0.0007	0.9970	0.0007
tae	0.3131	0.4773	0.3070	0.4807	0.2860	0.4869
texture	0.8159	0.1107	0.8247	0.1051	0.8006	0.1186
tumor	0.2696	0.6836	0.2674	0.6852	0.2520	0.6925
wine	0.7617	0.1012	0.7703	0.0964	0.7384	0.1071
zoo	0.8015	0.1079	0.8018	0.1075	0.7787	0.1144

由图 2 的数据点云分布和图 3 的点分布可以看出,RCRF 算法虽然在个体分类器上的平均误差略逊色于 RF 和 CRF

算法,但其在多样性方面却明显优于 RF 和 CRF 算法。

Breiman 指出随机森林的性能主要取决于单个分类器的精度和分类器间的多样性两个方面^[1]。从 κ -误差图的分析可以看出,RCRF 算法通过引入类别随机化机制使得个体分类器之间的多样性程度显著增加,而对单个分类器精度的影响并不明显,从而使得集成分类器的整体性能得到有效提高。同时,在准确率、F1-measure、Kappa 系数 3 种评价指标上对不同算法的实验分析结果也验证了 RCRF 算法的有效性,说明其在处理多分类问题时具有明显的优势。

结束语 分类方法是目前数据挖掘和机器学习领域的一个研究热点。随机森林因其良好的性能表现成为了常用的分类方法。为了提高随机森林在处理多分类问题上的性能表现,本文在充分考虑基分类器和类别之间关联性的基础上,提出了一种基于类别随机化的随机森林算法(RCRF)。从类别的角度出发,在随机森林原有的两种随机化的基础上增加了类别的随机化,针对不同类别,设计具有不同侧重点的基分类器,由于不同的分类器侧重区分的类别不同,所生成的决策树的结构也大不相同,这样就进一步增大了各个基分类器之间的差异性。最后,通过公开数据集上的实验分析验证了所提算法的有效性。实验结果表明,本文所提算法在解决多分类问题方面具有明显的优势。下一步我们将在保证基分类器多样性的同时进一步降低基分类器的误差率,使随机森林的性能得到全面的提升。

参 考 文 献

- [1] BREIMAN L. Random Forests [J]. *Machine Learning*, 2001, 45(1): 5-23.
- [2] FERNANDEZ-DELGADO M, CERNADAS E, BARRO S, et al. Do we need hundreds of classifiers to solve real world classification problems [J]. *Journal of Machine Learning Research*, 2014, 15(1): 3133-3181.
- [3] MEHER P K, SAHU T K, RAO A R. Identification of species based on DNA barcode using k -mer feature vector and random forest classifier [J]. *Gene*, 2016, 592(2): 316-324.
- [4] JOG A, CARASS A, ROY S, et al. Random forest regression for magnetic resonance image synthesis [J]. *Medical Image Analysis*, 2017, 35: 475-488.
- [5] WANG S, LIU J, BI Y Y, et al. Automatic recognition of breast gland based on two-step clustering and random forest [J]. *Computer Science*, 2018, 45(3): 247-252. (in Chinese)
王帅,刘娟,毕姚姚,等.基于两步聚类 and 随机森林的乳腺腺管自动识别方法 [J]. *计算机科学*, 2018, 45(3): 247-252.
- [6] FANELLI G, DANTONE M, GALL J, et al. Random forests for real time 3D face analysis [J]. *International Journal of Computer Vision*, 2013, 101(3): 437-458.
- [7] GALL J, YAO A, RAZAVI N, et al. Hough forests for object detection, tracking, and action recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(11): 2188-2202.
- [8] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees [J]. *Machine Learning*, 2006, 63(1): 3-42.
- [9] RODRIGUEZ J J, KUNCHEVA L I, ALONSO C J. Rotation forest: a new classifier ensemble method [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2006, 28(10): 1619-1630.
- [10] ZHANG L, SUGANTHAN P N. Random forests with ensemble of feature spaces [J]. *Pattern Recognition*, 2014, 47(10): 3429-3437.
- [11] ABELLÁN J, MANTAS C J, CASTELLANO J G. A random forest approach using imprecise probabilities [J]. *Knowledge-Based Systems*, 2017, 134: 72-84.
- [12] WANG Y, XIA S T, TANG Q, et al. A novel consistent random forest framework: bernoulli random forests [J]. *IEEE Transactions on Neural Networks & Learning Systems*, 2018, 29(8): 3510-3523.
- [13] YE Y, WU Q, HUANG J Z, et al. Stratified sampling for feature subspace selection in random forests for high dimensional data [J]. *Pattern Recognition*, 2013, 46(3): 769-787.
- [14] XIA J, LI L, LI L, et al. Adjusted weight voting algorithm for random forests in handling missing values [J]. *Pattern Recognition*, 2017, 69(C): 52-60.
- [15] HU C, CHEN Y, HU L, et al. A novel random forests based class incremental learning method for activity recognition [J]. *Pattern Recognition*, 2018, 78: 277-290.
- [16] BREIMAN L. Bagging predictors [J]. *Machine Learning*, 1996, 24(2): 123-140.
- [17] HO T K. The random subspace method for constructing decision forests [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(8): 832-844.
- [18] DEMSAR J. Statistical comparisons of classifiers over multiple data sets [J]. *Journal of Machine Learning Research*, 2006, 7(1): 1-30.
- [19] MARGINEANTU D D, DIETTERICH T G. Pruning adaptive boosting [C]// Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 1997: 211-218.