计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：　粒向量驱动的随机森林分类算法研究
作者：　张锟滨，陈玉明，吴克寿，侯贤宇
网络首发日期：　2023-09-04
引用格式：　张锟滨，陈玉明，吴克寿，侯贤宇．粒向量驱动的随机森林分类算法研究[J/OL]．计算机工程与应用.
https://link.cnki.net/urlid/11.2127.TP.20230904.1118.002

网络首发时间：2023-09-04 15:14:26
网络首发地址：https://link.cnki.net/urlid/11.2127.TP.20230904.1118.002

粒向量驱动的随机森林分类算法研究

张锟滨， 陈玉明， 吴克寿， 侯贤宇

厦门理工学院 计算机与信息工程学院，福建 厦门 361024

摘 要：粒计算是一种符合人类认知特性的计算范式,能够有效处理复杂数据.随机森林通过集成多个决策树来降低单个决策树的过拟合风险,但仍存在一定的过拟合问题.为了减少过拟合并提高分类性能,本研究在随机森林中引入了粒向量表示,提出了基于粒向量的随机森林分类算法.粒向量具有表示高维特征的能力,可以捕捉更多的数据模式;参照样本选择的随机性有助于控制过拟合;不同决策树使用不同的粒向量可以增加模型的多样性.实验结果表明,与传统随机森林以及其他改进算法相比,基于粒向量表示的随机森林算法具有较好的泛化能力,有效提高了分类的准确率,表明了基于粒向量的随机森林分类算法的正确性与有效性.

关键词：粒计算；粒向量；随机森林；集成学习
文献标志码:A 中图分类号:TP181 doi：10.3778/j.issn.1002-8331.2305-0204

Research on Granule Vectors Random Forest Classification Algorithm

ZHANG Kunbin, CHEN Yuming, WU Keshou, HOU Xianyu

College of Computer and Information Engineering, Xiamen University of Technology, Xiamen, Fujian 361024, China

Abstract：Granular computing is a computational paradigm that aligns with human cognitive characteristics, enabling the effective processing of complex data. Random Forest reduces the risk of overfitting associated with individual decision trees by ensembling multiple trees. However, it still faces some overfitting issues. To mitigate overfitting and enhance classification performance, this research introduces the concept of granular vector representation into Random Forest. Granular vectors possess the ability to represent high-dimensional features, thereby capturing more data patterns. The randomness in sample selection aids in controlling overfitting, while using different granular vectors for distinct decision trees enhances model diversity. Experimental results demonstrate that, compared to traditional Random Forest and other enhanced algorithms, the Random Forest algorithm based on granular vector representation exhibits superior generalization capabilities and significantly improves classification accuracy. This confirms the correctness and effectiveness of the granular vector-based Random Forest classification algorithm.

Key words：granular computing; granule vector; random forest; ensemble learning

基金项目：国家自然科学基金项目(61976183)。
作者简介：张锟滨(1997－),男,硕士生,研究方向为粒计算、机器学习,E-mail:337110638@qq.com;陈玉明(1977−),男,教授,博士生导师;吴克寿(1975−),男,教授,博士生导师;侯贤宇(1997－),男,硕士生。

1979 年,美国科学家 Zadeh 首次提出并讨论了模糊信息粒度化问题[1].这一概念的提出,引发了不同领域的学者对信息粒化的探索与研究.1988 年,Lin 提出了邻域系统并研究了其与关系数据库的关系[2],1996 年,Lin 第一次提出了粒计算(granular computing)的概念,他给出了信息处理中一种新的概念与计算范式,并在数据挖掘领域进行应用实践[3].在 Lin 的研究基础上,Yao 定义了一种邻域关系[4],进而提出了邻域粒计算[5],并将其应用于数据挖掘等领域.2000 年后,随着粒计算热度不断提高,国内学者也加大了对粒计算的研究力度.苗夺谦等人对知识的粒计算进行研究,给出了属性重要度启发式的属性最小约简算法,及基于协调度的决策树构造方法[6].胡清华等人分析了邻域的约简,在文献[7]中提出了一种基于邻域关系的粒化方式,从而实现了实数空间中的粒计算,并在此基础上设计了邻域分类器[8-9].Chen 在文献[10-12]中提出了基于单特征模糊粒化结合卷积的分类模型和基于信息粒的随机模糊粒度决策树算法,将模糊粒化与机器学习算法结合,进行聚类与分类,并分析了粒的不确定性和距离度量.从信息粒度的角度分析,不难发现聚类和分类有很大的相通之处:聚类是在一个统一的粒度下进行计算,而分类是在不同的粒度之下进行计算[13-14].粒和粒化是符合人类认知特性的范式,在大数据、数据挖掘以及复杂数据建模中有着重要作用,并广泛应用于诸多领域[15-17].

随机森林(Random Forest, RF)[18]是一种集成分类算法,其核心思想是通过建立多个决策树来降低单个决策树的过拟合风险.每个决策树都是在不同的样本和特征集上训练,这种随机性可以减少算法的方差,并提高模型的泛化能力.这些决策树可以并行训练.在随机森林中,每个决策树的输出被视为一个投票.在分类问题中,随机森林会将实例分配给获得最多投票的类别,具有高稳定性、模型泛化能力强,易并行化等优点,并且由于其在分类任务上相比于其他算法具有更好的表现,因此广泛应用于检测系统[19]、推荐系统[20]、诊断系统[21].随机森林的起始性能往往比较差,特别是只有一个基学习器时,这是因为基学习器的训练过程中加入了属性扰动,导致基学习器的性能降低[22].但是,随着基学习器的个数增加,随机森林产生的集成学习器的性能会得到很大的提升,即最终泛化误差会收敛到最小.根据文献[18],当树的数目足够大时,随机森林的泛化误差的上界收敛于下面的表达式:

$$\text{泛化误差} \le \frac{\bar{\rho}(1-s^2)}{s^2} \tag{1}$$

其中$\bar{\rho}$是树之间的平均相关系数,s是度量树型分类器强度的量.通过分析式(1)可知,随机森林的过拟合风险可以通过 Bagging 和特征随机选择来控制,但仍然存在一定的过拟合风险.相关性的存在使得随机森林的泛化误差略高于独立决策树的误差.此外,决策树和随机森林本身也有一定的偏差,特别是在复杂模式或特定样本分布的情况下.针对以上问题,本文在随机森林分类算法中引入粒向量,提出了基于粒向量的随机森林分类算法,该算法主要有以下优势:

(1)高维特征表示:粒向量引入了高维特征表示,将数据点映射到一个更大的特征空间.这有助于捕捉更多的数据关系和模式,尤其在处理复杂的非线性关系时效果更好.

(2)参照样本选择的随机性:随机森林算法在每棵决策树构建时随机选择特征,而粒向量每个维度都对应多个随机选择的参照样本特征.这种随机性有助于减少过拟合,增加模型的泛化能力.

(3)模型多样性:随机森林通过集成多棵决策树来进行预测,每棵决策树都是使用不同的数据子集和特征子集构建的.引入粒向量后,每棵决策树的特征子集也是随机选择的.这样可以增加模型的多样性,减少模型的方差,提高模型的鲁棒性.

在下文将首先详细介绍粒向量的定义和算法,以及其在随机森林中的应用.随后,提出基于粒向量的随机森林分类算法.最后,使用 UCI 数据集对基于粒向量的随机森林分类算法与传统随机森林算法和其他方法进行性能对比,验证粒向量算法的正确性和有效性,为随机森林算法的优化探索了一个新方向.

1 相关工作

1.1 粒子与粒向量

传统随机森林算法的输入对象为样本,在粒计算理论中,输入则为一个由粒子组成的粒向量.文献[23]提出了粒的构造方法,可在列(属性)上进行粒化;文献[17]提出了在单特征上粒化为粒子,多特征上粒化构造粒向量的具体方法,并进一步给出了粒的结构、距离度量等定义.

设数据集为$U=(X\cup P,C)$,其中$X=\{x_1,x_2,...,x_n\}$为训练样本集;$P=\{p_1,p_2,...,p_k\}\subseteq X$为随机抽取的局部样本作为粒化参照样本;$m$维特征集合为$C=\{c_1,c_2,...,c_m\}$.给定单样本$x\in X$,对于单特征

$c\in C$，$v(x,c)\in[0,1]$表示样本x在特征c上归一化后的值.则x与p在单特征c上的相似度为:

$$s_c(x,p)=1-|v(x,c)-v(p,c)|. \tag{2}$$

定义 1 给定数据集$U=(X\cup P,C)$,对于任一样本$x\in X$和参照样本集$P=\{p_1,p_2,...,p_k\}$,以及任一单特征$c\in C$,则x在参照样本p中的特征c上进行粒化,形成的粒子定义为:

$$g_c(x)=\{g_c(x)_j\}_{j=1}^{k}=\{r_j\}_{j=1}^{k}=\{r_1,r_2,...,r_k\},$$

其中$r_j=s_c(x,p_j)$表示样本x以p_j为参考,在单特征c上的相似度.易知$s_c(x,p_j)\in[0,1]$,因此$r_j\in[0,1]$.粒子由粒核组成,$g_c(x)$称为粒子,则$g_c(x)_j$称为第j^{th}个粒核.若$\forall r_j=1$,则为1-粒子,简写为1;若$\forall r_j=0$,则为0-粒子,简写为0.

定义 2 设为数据集$U=(X\cup P,C)$,对于任一样本$x\in X$,任一特征子集$A\subseteq C$,设$A=\{a_1,a_2,...,a_m\}$,则在特征子集A上的粒向量x定义为:

$$G_A(x)=(g_{a_1}(x),g_{a_2}(x),...,g_{a_m}(x))^T,$$

其中$g_{a_m}(x)$是样本x在特征a_m上的粒子.为方便计,特征集$A=\{a_1,a_2,...,a_m\}$用整数标记,则粒向量表示为$G_A(x)=(g_1(x),g_2(x),...,g_m(x))^T$.

粒向量由粒子组成,粒子又由粒核构成.因此,粒向量可以是一个粒核矩阵的形式,表示为:

$$G(x)=\begin{bmatrix} g_1(x)_1, & g_1(x)_2, & \cdots, & g_1(x)_k \\ g_2(x)_1, & g_2(x)_2, & \cdots, & g_2(x)_k, \\ \vdots & \vdots & \ddots & \vdots \\ g_m(x)_1, & g_m(x)_2, & \cdots, & g_m(x)_k \end{bmatrix}$$

$$=\begin{bmatrix} r_{11}, & r_{12}, & \cdots, & r_{1k} \\ r_{21}, & r_{22}, & \cdots, & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1}, & r_{m2}, & \cdots, & r_{mk} \end{bmatrix},$$

与原数据集相比,粒核矩阵的大小受参照样本数量的影响:参照样本越多,粒核矩阵越大;参照样本越小,则粒核矩阵越小.粒向量也可以用另外一种形式表示为:

$$G_A(x)=(g(x)_1,g(x)_2,...,g(x)_k),$$

其中$g(x)_j=(g_1(x)_j,g_2(x)_j,...,g_m(x)_j)^T$.粒向量由粒子组成,而粒子是一个集合的形式.因此,粒向量的元素是集合,与传统向量不一样,传统向量的元素是一个实数.

1.2 粒的运算

上节主要阐述随机抽取部分样本作为参照样本,然后对训练集样本在参照样本中进行粒化后,构造出粒子与粒向量.这一小节定义粒的相关运算与距离度量,建立基于粒向量的随机森林运算基础.

定义 3 设粒子为$g_c(x)=\{r_j\}_{j=1}^{k}$,其大小定义为:

$$|g_c(x)|=\sum_{j=1}^{k}r_j, \tag{3}$$

由粒子的定义可知$r_j\in[0,1]$,因此$0\le|g_c(x)|\le k$.

定义 4 设$g_a(x)=\{s_j\}_{j=1}^{k}$,$g_b(x)=\{t_j\}_{j=1}^{k}$为样本x在特征a,b上的两个粒子,则两个粒子的加、减、乘、除运算定义为:

$$g_a(x)+g_b(x)=\{s_j+t_j\}_{j=1}^{k};$$

$$g_a(x)-g_b(x)=\{s_j-t_j\}_{j=1}^{k};$$

$$g_a(x)*g_b(x)=\{s_j*t_j\}_{j=1}^{k};$$

$$g_a(x)/g_b(x)=\{s_j/t_j\}_{j=1}^{k}.$$

定义 5 设$g_a(x)=\{s_j\}_{j=1}^{k}$,$g_a(y)=\{t_j\}_{j=1}^{k}$为样本x,y在特征a上的两个粒子,则两个粒子的加、减、乘、除运算定义为:

$$g_a(x)+g_a(y)=\{s_j+t_j\}_{j=1}^{k};$$

$$g_a(x)-g_a(y)=\{s_j-t_j\}_{j=1}^{k};$$

$$g_a(x)*g_a(y)=\{s_j*t_j\}_{j=1}^{k};$$

$$g_a(x)/g_a(y)=\{s_j/t_j\}_{j=1}^{k}.$$

两个粒子的加减乘除运算结果为一个粒子.定义4是针对同一个样本在不同特征集合上粒化后不同粒子的运算,而定义5则是应用在不同样本在同一特征集合上粒化后粒子上的运算.

定义 6 设$g_c(x)=\{s_j\}_{j=1}^{k}$和$g_c(y)=\{t_j\}_{j=1}^{k}$为两个粒子,则粒子的欧氏距离度量为:

$$o(g_c(x),g_c(y))=\sqrt{\sum_{j=1}^{k}\{s_j-t_j\}^2}.$$

定义 7 设$G_C(x)=(g_1(x),g_2(x),...,g_m(x))^T$,

$G_C(y)=(g_1(y),g_2(y),...,g_m(y))^T$ 为两个粒向量,则粒向量的欧氏距离度量为:

$$o(G_C(x),G_C(y))=\| G_C(x)-G_C(y)\| =\sqrt{\sum_{i=1}^{m}(o(g_i(x),g_i(y)))^2},$$

其中 $o(g_i(x),g_i(y))$ 为粒子的欧氏距离.

1.3 粒范数

这一小节进一步定义粒范数.粒范数可用于衡量特征的重要性和稀疏性.通过引入粒范数作为正则化项,可以促使模型选择具有较大权重的特征,同时抑制那些具有较小权重或冗余的特征.这有助于降低模型的复杂性,避免过拟合,并提高泛化能力.

定义 8 设 $g=\{r_j\}_{j=1}^{n}$ 为粒子,则粒子的范数定义为:

(1) 粒子-1范数:

$$Norm-1(g)=\|g\|_1=\sum_{j=1}^{n}|r_j|; \tag{4}$$

(2) 粒子-2范数:

$$Norm-2(g)=\|g\|_2=\sqrt{\sum_{j=1}^{n}r_j^2}; \tag{5}$$

(3) 粒子-p范数:

$$Norm-p(g)=\|g\|_p=(\sum_{j=1}^{n}r_j^p)^{\frac{1}{p}}; \tag{6}$$

(4) 粒子-max范数:

$$max(g)=\|g\|_{max}=\max_{1\le j\le n}\{\|r_j\|\}; \tag{7}$$

(5) 粒子-min范数:

$$min(g)=\|g\|_{min}=\min_{1\le j\le n}\{\|r_j\|\}. \tag{8}$$

定义 9 设m维粒向量为 $G=(g_1,g_2,...,g_m)=(g_i)_{i=1}^{m}$,则粒向量范粒子定义为:

(1) 粒向量-1范粒子:

$$\|G\|_1=\sum_{i=1}^{m}g_i; \tag{9}$$

(2) 粒向量-2范粒子:

$$\|G\|_2=\sqrt{\sum_{i=1}^{m}g_i*g_i}=\sqrt{G\cdot G}; \tag{10}$$

(3) 粒向量-p范粒子:

$$\|G\|_p=(\sum_{i=1}^{m}g_i^p)^{\frac{1}{p}}. \tag{11}$$

粒向量的范粒子运算结果为粒子,提供了一条由粒向量转化为粒子的途径.

定义10 设 m 维粒向量为 $G=(g_1,g_2,...,g_m)=(g_i)_{i=1}^{m}$,则粒向量的范数定义为:

(1) 粒向量-11范数

$$\|G\|_{11}=\sum_{i=1}^{m}\|g_i\|_1=\sum_{i=1}^{m}\sum_{j=1}^{n}|r_{ij}|; \tag{12}$$

(2) 粒向量-12范数

$$\|G\|_{12}=\sum_{i=1}^{m}\|g_i\|_2=\sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n}r_{ij}^2}; \tag{13}$$

(3) 粒向量-21范数

$$\|G\|_{21}=\sum_{i=1}^{m}\|g_i\|_1^2=\sqrt{\sum_{i=1}^{m}(\sum_{j=1}^{n}|r_{ij}|)^2}; \tag{14}$$

(4) 粒向量-22范数

$$\|G\|_{22}=\sqrt{\sum_{i=1}^{m}\|g_i\|_2^2}=\sqrt{\sum_{i=1}^{m}(\sqrt{\sum_{j=1}^{n}r_{ij}^2})^2} =\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}r_{ij}^2} \tag{15}$$

粒向量的范数运算结果为实数,粒子的范数运算结果也为实数,这些运算提供了粒向量与粒子转化为实数的途径.

2 基于粒向量的随机森林算法

基于粒向量的随机森林分类算法是有监督分类算法,它结合粒计算理论以及随机森林思想,将可并行的粒与集成学习融合,对多特征描述下的粒向量进行分类,以提高随机森林的性能.为了设计基于粒向量的随机森林分类算法,需先定义基于粒向量的随机森林结构,阐述基于粒向量的随机森林分类算法的原理.

2.1 基于粒向量的随机森林原理

根据定义 1 和定义 2,数据集将以粒矩阵的形式输入随机森林.经相似度粒化的随机森林算法随机选出的粒向量和粒子,参照文献[16]的思想构造粒决策树.本文根据公式(2)进行相似度粒化,原数据通过粒化生成的粒向量以局部参照样本进行粒化,可以通过局部信息构造相关系数较低的相似度粒核矩阵;在所有样本上进行粒化,所以算法能够把握全局信息进行决策,进而能够有效提高算法的准确率.通过公式(1)分析传统随机森林存在的问题,本文提出的基于粒向量的随机森林算法具有以下优势:

(1)GvRF 可构造的决策树数量是 RF 算法中的 $|g(x)|$ 倍,能快速提高基学习器数量,以提高随机森林的收敛速度;

(2)由于参照样本的选取具有随机性,生成的粒矩阵能够提供多个相关性较弱的分类器,能有效降低相

关系数$\overline{\rho}$.

(3)用于构建粒向量的参照样本均来自原始数据,通过随机选取可以更好拟合原始数据分布,以提高算法在复杂模式或不同分布的数据集中的性能.

2.2 基于粒向量的随机森林模型结构

参考随机森林的结构,基于粒向量的随机森林模型分为五个部分:输入层、粒化层、抽样层、并行层、决策层、输出层,其模型结构如图 1 所示.

首先,输入信息空间$IS=(U,F)$,其中样本集为$U=\{x_1,x_2,\ldots,x_n\}$,属性集为$F=\{f_1,f_2,\ldots,f_m\}$,对样本集进行归一化操作.GvRF 模型的粒化层随机选取参照样本构成参照样本集$P=\{p_1,p_2,...,p_k\}$,并使用公式(2)在所有属性下进行粒化,将原数据集粒化成为一个粒核矩阵$GT=\{G(x_1),G(x_2),\ldots,G(x_n)\}$,粒核矩阵的大小由参照样本的多少决定.粒化过后的相似粒矩阵GT通过随机抽取粒向量用于构造决策树根节点的训练数据,随机抽取粒子进行节点的分裂.在并行层,粒核矩阵GT将被处理成多个新的粒核矩阵$\{GT_j\}_{j=1}^{k}=\{GT_1,GT_2,\ldots,GT_k\}$,其中$k$为样本的个数.每个粒核矩阵用于构造粒决策树,构造好的粒决策树可进行并行运算.最后通过预测层得出每棵树决策的类别,形成决策集,最后通过投票在输出层确定该样本的输出类别.

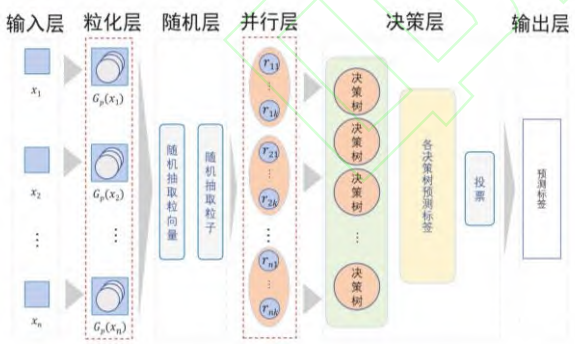

图 1 基于粒向量的随机森林模型结构

Fig.1 Granule vector based random forest model structure

2.3 基于粒向量的随机森林算法流程

上一小节主要分层具体描述 GvRF 算法模型结构.这一小节主要阐述基于粒向量的随机森林算法流程.

算法 1 基于粒向量的随机森林算法

输入: 信息空间$IS=(U,F,T)$,其中样本集为$U=\{x_1,x_2,\ldots,x_n\}$,对应类别标签$T=\{y_1,y_2,\ldots,y_n\}$,属性集$F=\{f_1,f_2,\ldots,f_m\}$;参照集大小k;基学习器的个数N.

输出: 预测类别标签$C=\{y_1',y_2',\ldots,y_n'\}$.

步骤 1 将样本集U进行数据结构化和归一化处理,记为U';

步骤 2 随机选取k个样本组成粒化参照样本集$P=\{p_1,p_2,\ldots,p_k\}$;

步骤 3 预处理后的样本集U'在参照集P中进行粒化,形成粒核矩阵$GT=\{G(x_1),G(x_2),\ldots,G(x_n)\}$,其中$G(x_i)=(g_1(x_i),g_2(x_i),\ldots,g_m(x_i)),i\in[1,n]$;

步骤 4: 按照 25%的比例将粒核矩阵与对应类别标签$T=\{y_1,y_2,\ldots,y_n\}$对应并进行随机划分,生成训练集GT_{train}和测试集GT_{test};

步骤 5: for $t=1$ to N do

步骤 6: 将训练集GT_{train}作为随机森林算法的输入进行随机采样得到GT'_{train};

步骤 7: 对GT'_{train}进行并行化处理,生成k倍的基学习器,并以真实标签为目标进行训练;

步骤 8: end for

步骤 9: 组合所有的由粒核矩阵训练的基学习器构成基于粒向量的随机森林$Forest_G$

步骤 10: for $model_G$ in $Forest_G$ do

步骤 11: 输入测试集GT_{test},得到由$model_{G_i}$预测的一组标签c_i;

步骤 12: end for

步骤 13: 组合每个基学习器得到的预测标签c_i进行投票得到一组预测类别标签$C=\{y_1',y_2',\ldots,y_n'\}$;

步骤 14: 输出预测类别标签 $C=\{y_1',y_2',\ldots,y_n'\}$.

算法结束.

根据算法 1,GvRF 算法中N和k共同决定了随机森林中基学习器的数量.与传统随机森林算法相同,基于粒向量的随机森林算法的时间复杂度主要包括基学习器的训练和预测阶段.在训练阶段,需要构建多个决策树.每个决策树的构建时间复杂度通常为$O(m\sqrt{n})$,其中m是属性数量,n是样本数量.粒具有可并行化的特性,对于基学习器的个数N和参照集大小k,算法采用并行化处理,总体时间复杂度约为$O(Nm\sqrt{n})$.在预测阶段,随机森林中的每棵决策树都需要遍历,时间复杂度为$O(N\sqrt{n})$.因此,GvRF 算法的总体时间复杂度约为$O(Nm\sqrt{n})$.对于每棵决策树,存储的空间复杂度为$O(m)$.而基学习器的个数N和参照集大小k也会增加存储开销.因此,GvRF 算法的总体空间复杂度约

为$O(Nkm)$.通过以上分析,GvRF 算法相比于传统随机森林算法而言,由于其粒的特性,在增加模型输入信息的同时,没有增加模型的时间复杂度,这也是将粒向量引入随机森林算法的优势之一.

传统随机森林算法中每个基学习器只使用部分样本和特征进行训练.这样可以增加样本和特征之间的差异性,减少模型对于训练集的过拟合.通过集成多个基学习器的预测结果,可以降低方差并提高模型的稳定性和泛化能力.由于引入了参照集选择的随机性,GvRF 算法在传统随机森林算法的基础上,进一步提供了多个高差异性的基学习器,这使得 GvRF 算法相比于传统随机森林算法具有快速收敛的性质,同时也进一步提高了模型的泛化能力.

分析算法 1 可知,相比于传统随机森林算法,GvRF 算法需要额外指定参照集大小k,即算法所需的超参数有:基学习器的个数N和参照集大小k.它们共同决定了模型的学习器大小,以及算法时间与储存开销.参数的变化对算法的分类结果和运行性能有很大的影响,因此需要选择合适的数值.通过实验结果表明,综合考虑分类效果以及算法成本,GvRF 算法中基学习器的个数N的合理取值范围为 0~50,参照集大小k的合理取值范围为 4~10.

3 实验与分析

为验证所设计基于粒向量的随机森林算法(GvRF)的综合有效性,本文采用了 UCI 中的多个高维小样本数据集进行实验,所有数据集的描述性信息如表 1 所示.

表 1 实验采用的 UCI 数据集

Table 1 UCI dataset used in the experiment

数据集名称	样本数(个)	特征数(个)	类别数(个)
Wine	178	13	3
Seed	209	7	3
Glass	213	9	6
Heart	269	13	2
Iris	150	4	3
Column3C	310	6	3
Diabetes	768	8	2
Breast cancer	569	30	2

由于每个数据集中特征量纲不同,所以需要对每个数据集进行最大最小值归一化处理,将每个特征数据变换到$[0,1]$的区间之中.归一化公式如下:

$$X_{norm}=\frac{X-X_{min}}{X_{max}-X_{min}} \tag{16}$$

预处理结束后,对预处理的数据进行相似度粒化,单特征上形成粒子,多特征上形成粒向量,GvRF 算法的输入即为多个粒向量组成的相似粒矩阵.本文使用 GvRF 算法与 RF 算法在 UCI 数据集上,将算法分类效果作为评价指标进行对比.本文还讨论了两个超参数:基学习器个数N,参照集大小k对提出算法的影响.

3.1 GvRF 与 RF 对比实验

对于表 1 中的 8 个 UCI 数据集,实验采用提出的基于粒向量的随机森林算法(GvRF)和随机森林算法(RF)进行分类效果的比较.对比实验中包含准确率、召回率和 F1 分数三种评估指标,每个指标的值都是均值(Mean%)加上标准差(Std%).数据集首先通过公式(16)进行归一化,输入 RF 算法的为归一化后的数据,输入 GvRF 算法的数据则还需要经过公式(2)的粒化操作转换成粒核矩阵.本次实验的超参数设置如下:基学习器数量N设定为 25,参照集大小k设定为 5,其他条件均保持一致,所有实验均采用十折交叉验证.结果如表 2 所示.可以看出,在 7 个数据集上,GvRF 方法在准确率、召回率和 F1 分数三个评价指标上均优于 RF 方法.具体来看,在准确率方面,GvRF 方法的提高范围在 0.56%到 2.62%之间.在召回率方面,GvRF 方法的提高范围在 1.92%到 3.86%之间,平均提高 2.80%.在 F1 分数方面,GvRF 方法的提高范围在 1.54%到 3.35%之间,平均提高 2.34%.这表明 GvRF 对比于 RF 在提高模型分类性能的广度和深度上都取得了较好效果.

表 2 GvRF 与 RF 在不同数据集中的性能对比 (Mean%±Std%)

Table 2 Performance comparison of GvRF and RF across different datasets (Mean%±Std%)

	Wine		Seed		Glass		Heart	
	RF	GvRF	RF	GvRF	RF	GvRF	RF	GvRF
准确率	98.33±2.55	**98.89±2.22**	92.79±5.83	**93.29±7.15**	78.81±9.32	**80.26±8.60**	80.66±7.59	**83.28±7.44**
召回率	98.33±2.69	**99.00±2.13**	91.99±6.69	**92.57±8.04**	70.47±15.45	**74.03±12.61**	80.02±8.35	**82.85±5.82**
F1 分数	98.23±0.51	**98.86±2.28**	92.49±6.15	**92.64±8.06**	68.43±15.97	**73.97±13.68**	79.64±8.45	**82.68±7.46**
	Iris		Column3C		Diabetes		Breast cancer	
	RF	GvRF	RF	GvRF	RF	GvRF	RF	GvRF
准确率	95.33±4.27	95.33±6.00	84.19±6.20	**84.84±5.41**	75.26±5.06	**76.82±3.43**	95.96±1.93	**96.31±2.54**
召回率	95.46±4.34	**95.89±5.09**	79.78±8.55	**80.89±6.43**	71.05±4.48	**73.06±3.21**	94.96±2.99	**95.91±2.96**
F1 分数	94.95±4.60	**94.96±6.63**	78.65±9.25	**79.53±7.69**	71.37±4.69	**73.22±3.20**	95.41±2.39	**96.00±2.76**

但是,GvRF 方法的提高幅度在不同数据集的评价指标之间也存在差异.例如,GvRF 与 RF 的提高幅度在数据集 Heart 与数据集 Iris 上存在明显差异.在数据集 Heart 上,GvRF 方法的召回率提高 2.87%,F1 分数提高 3.04%,而在数据集 Iris 上,这两个指标的提高幅度仅为 0.43%和 0.01%.这表明 GvRF 方法在高维数据集上表现出更强的优势,这主要是因为高维数据集可以结合相似度粒化方法提供更丰富的信息以供其进行决策.

综上,表格结果显示 GvRF 方法相比 RF 方法在高维小样本数据分类性能上获得了较为全面和稳定的提高.同时,也应注意到算法性能的提高在不同数据集和评价指标之间的差异,这需要在算法比较和选择时综合考虑其他因素包括参照样本数量 k 、基学习器数量 N 以及对不同数据集采用不同的策略,以做出更加准确的决策.

3.2 参数的影响

对于 GvRF 和 RF 算法,不同基学习器的数量同样影响算法的分类效果.本文提出的 GvRF 算法主要由基学习器数量 N 和参照样本数量 k 两个超参数共同作用,本小节通过实验讨论这两个参数对 GvRF 算法的具体影响.

3.2.1 基学习器数量 N

为探索不同大小的基学习器数量 N 对算法的影响,本节在每个数据集上以不同的基学习器数量进行实验.实验以[2,100]为区间,2 为步长确定基学习器数量 N ,参照样本数量 k 为固定值 5 进行,其余条件均保持不变,每组实验均进行十折交叉验证.图 2 为 GvRF 在不同数据集中不同基学习器数量 N 的实验结果.

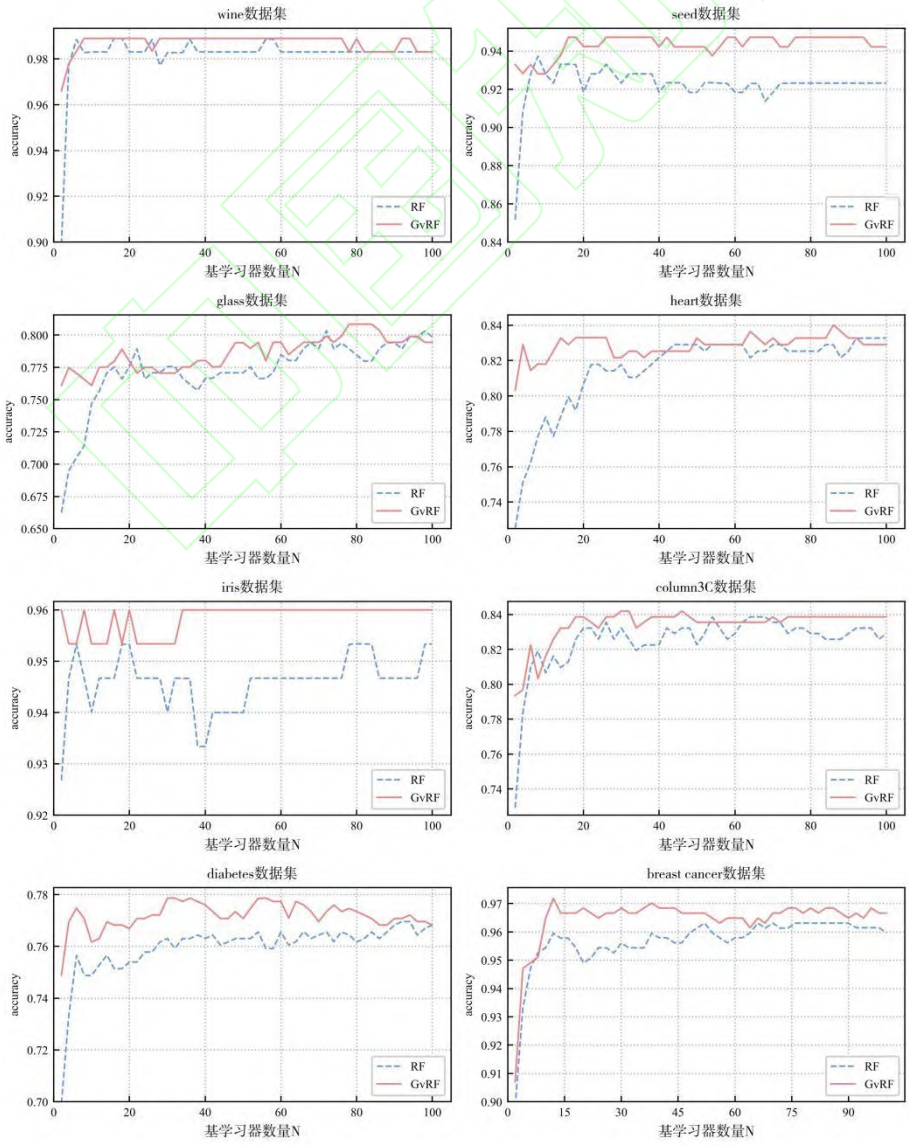

图 2 GvRF 在不同基学习器数量 N 的准确率

Fig.2 The accuracy of GvRF with different N

根据图 2 可知,在所有数据集的实验中,GvRF 对于 RF 算法在不同基学习器数量下准确率均有一定的提升.从收敛速度看,由于 GvRF 算法采用相似度粒化使数据集以粒向量形式扩充基学习器,其收敛速度在各数据集上均优于 RF 算法,尤其在 heart 数据集上较为明显,Iris 数据集由于样本数与特征数都相对较小,GvRF 算法最开始就处于最优值,并在之后小幅震荡.从收敛趋势来看,除了 glass 数据集仍然处于上升趋势,其他数据集均趋于收敛.可以观察到,多数情况下,GvRF 算法收敛结果要高于传统 RF 算法,但在基学习器数量 N 的值超过 50 后,部分数据集上的指标也出现了小幅下降的趋势,但总体指标仍然高于 RF.这说明基学习器数量 N 的变化并没有明显影响 GvRF 算法对于 RF 算法的性能提升.

3.2.2 参照样本数量 k

在不同参照样本数量的实验中,将基学习器数量 N 设定为固定值 25,参照样本数量设定在[1,20]区间内,步长为 1,其它条件均保持不变,每个实验均进行十折交叉验证,结果如图 3 所示.

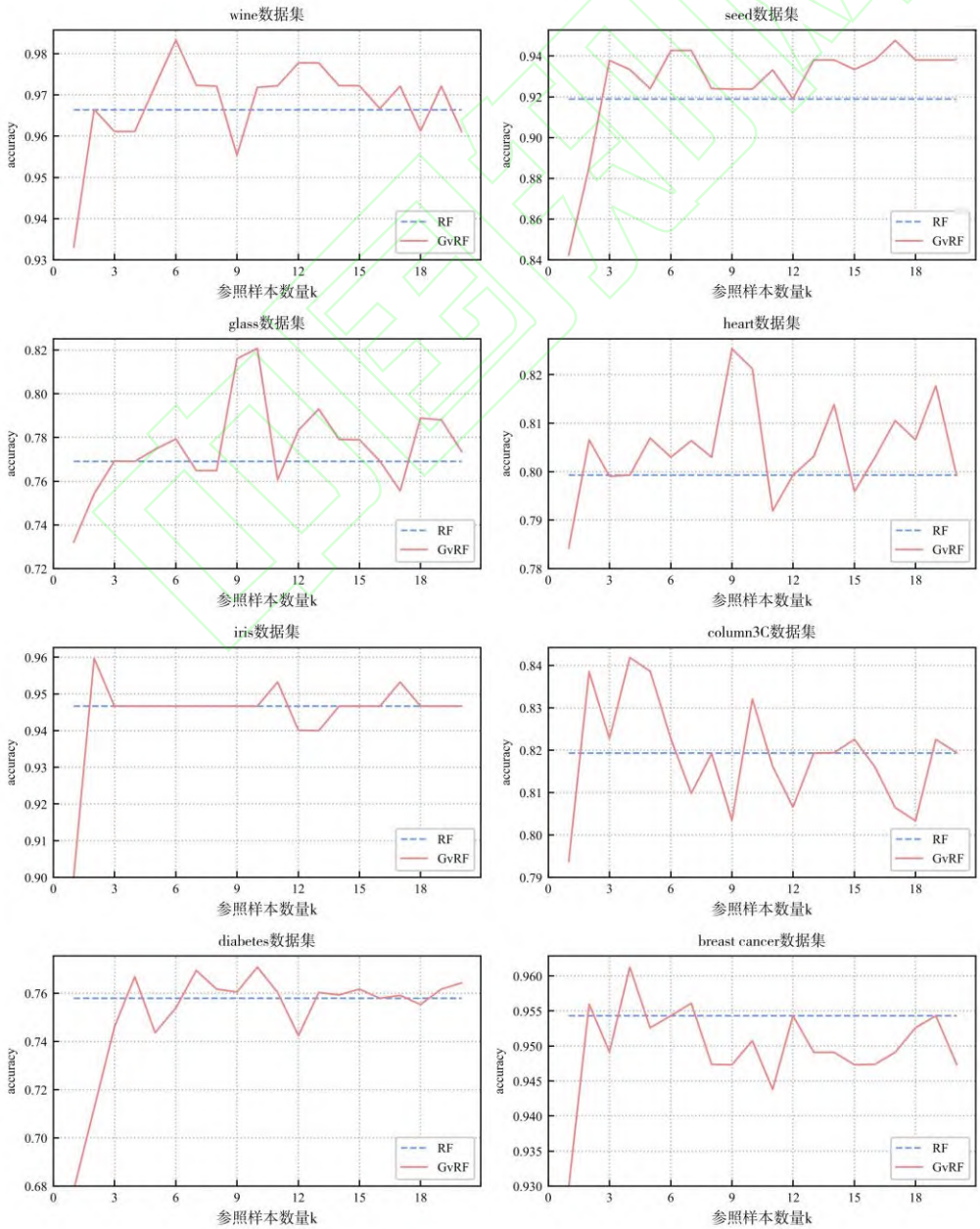

图 3 GvRF 在不同参照样本数量 k 的准确率

Fig.3 The accuracy of GvRF with different k

对图 3 分析可知,在所有数据集上,GvRF 相对于 RF 在准确率指标上均有出不同程度的提升,其中在 Glass 和 Heart 数据集上提升幅度尤为明显,在 Iris 和 Diabetes 上提升幅度较小,且不同数量的参照样本对同一数据的决策准确率有着较大幅度的影响.从趋势上分析,随着参照样本个数的不断提升,GvRF 算法性能在初期($k\in[4,10]$)可以快速提升,在出现峰值数据后,部分数据集例如 Seed 和 Diabetes 数据集的准确率趋于稳定,在其他数据集上的准确率呈下降趋势.

结合图 2 和图 3 实验内容可以看出,基学习器数量 N 和参照样本数量 k 两个超参数共同决定了 GvRF 算法的分类精度,并且由于参照样本的选择具有随机性,参数 k 对提出的算法具有更大的影响.根据泛化误差公式(5),本文提出的 GvRF 算法的优势在于:可以随机选择参照样本并通过相似度粒化的方式快速构造出多个相关系数较低的基学习器,在减少了泛化误差的同时提高了其收敛速度.值得注意的是,当 N 和 k 的值相对偏大时,GvRF 的性能出现下降的趋势,这个现象在变化参数样本数量 k 时尤为明显.综合以上分析,本文提出的 GvRF 算法在所实验的数据集中均有不同程度的提升,主要受到基学习器数量 N 和参照样本数量 k 两个超参数的影响.其中对于高维小样本数据的提升幅度更大,这充分说明了 GvRF 算法的正确性和有效性.考虑算法效率等因素,推荐在高维小样本数据集中,参数 k 的值选择较小的参数,例如[4,10].参数 N 的值则由于其收敛后具有较稳定的分类表现,可以根据不同数据集进行范围较大的自由选择.

3.3 GvRF 与其他方法对比实验

本小节主要对比了提出的基于粒向量的随机森林分类算法和以下对比算法:

(1)传统随机森林(Random Forest):建立多个决策树来降低单个决策树的过拟合风险.每个决策树都是在不同的样本和特征集上训练.

(2)极限随机树[24](Extra-Trees): 极限随机树是一种对传统随机森林的改进,其在构建决策树时,会随机选择特征和切分点,而不是使用最优的选择.不同样本实验中统一设置特征采样数为总特征数的平方根个特征.

(3)旋转森林[25](Rotation Forest): 旋转森林是一种利用特征旋转增加模型多样性的方法,每棵树都在经过特征旋转变换后的特征空间上构建,旋转变换通过主成分分析(PCA)等方法实现.不同样本实验中统一设置旋转次数为 3,随机旋转的角度范围.

(4)XGBoost[26](eXtreme Gradient Boosting): XGBoost 是一种基于梯度提升树的集成学习算法,在梯度提升树的基础上引入了正则化项,通过控制模型的复杂度来防止过拟合.不同样本实验中统一设置学习率为 0.1,采用 L2 正则化.

本次实验中,GvRF 将基学习器数量设置为 25,参照样本数量设置为 4.除了不同算法特有的超参数,决策树部分超参数统一设置如下:基学习器数量为 25,最小分割样本数为 2,最小叶子节点样本数为 1,树的最大深度为 3,分裂标准为基尼系数($gini$).表 3 比较了 GvRF 算法和其他 4 种算法:RF、ET(文献[24]方法)、RoF(文献[25]方法)、XGBoost(文献[26]方法)在 8 个不同数据集上的分类准确率.

表 3 GvRF 与其他算法对比实验(%)

Table 3 Accuracy comparison of GvRF and other algorithms on different datasets(%)

datasets	RF	ET[24]	RoF[25]	XGBoost[26]	GvRF
Wine	97.19	98.33	97.75	96.60	**98.89**
Seed	91.88	92.83	92.34	**93.33**	**93.33**
Glass	76.10	77.05	72.30	73.72	**77.46**
Heart	82.16	80.68	79.18	**83.26**	82.90
Iris	**96.00**	94.67	95.33	94.67	**96.00**
Column3C	81.93	82.58	78.70	82.25	**84.19**
Diabetes	75.78	74.61	72.78	**76.69**	**76.69**
Breast cancer	95.61	96.66	94.02	**96.67**	96.13

从表 3 数据可知,GvRF 算法在大多数数据集上表现较好,特别是在 Wine、Glass、Column3C 数据集上的准确率最高,分别为 98.89%,77.46%,84.19%.在 Seed 和 Diabetes 数据集上,GvRF 算法的准确率与 XGBoost 相当.在 Heart 数据集上,XGBoost 略优于 GvRF.另一方面,综合数据集数据,可以看出 GvRF 算法在小样本数据集(如 Iris)和大样本数据集(如 Diabetes)上表现都比较好,同时也能处理特征数相对较多的问题(如 Breast cancer),这表明 GvRF 算法具有较强的泛化能力,对样本量和特征数量较不敏感,能够较好地扩展到不同规模和结构的数据集.与 XGBoost 相比,GvRF 处理小样本数据集的能力更强, XGBoost 则在高维特征数据集上表现较好,因此 GvRF 更适合样本量不足的场景.综合来看,GvRF 算法相比其他算法有更好的泛化能力,能够在不同类型的数据集上都获得较高的分类准确率.

针为了进一步验证 GvRF 算法的泛化性能,我们分别以 Heart 和 Breast Cancer 两个数据集为例,比较了 GvRF 与其他算法在不同基学习器数量 N 下的分类准确率,如图 4 展示.结果表明:在 Heart 数据集中,当基学习器数量较小时,各算法的分类准确率较低,都在 0.80-

0.82 之间,但 GvRF 略高于其他算法.随着 N 的增加,所有算法的准确率均有提升.当 N 达到 100 时，GvRF 算法的准确率为 0.85,高于 XGBoost 与其他算法.这表明随着基学习器数量的增加,GvRF 算法的优势逐渐增加.在 Breast Cancer 数据集中,各算法的分类准确率维持在 0.94-0.96 的较高水平,GvRF 仅比 RF 略高 0.52 个百分点.随着 N 的增加,GvRF 的准确率稳步提升,在 $N=100$ 时 GvRF 的准确率高于传统随机森林与旋转森林，低于 XGBoost 算法.这也说明在此类数据集中,基学习器数量的增加对 GvRF 准确率提升具有一定的帮助.值得注意的是,提升 GvRF 的基学习器数量在提高准确率的同时也相应增加了算法开销，在具体应用时需要针对不同数据集进行参数寻优.综上,增加基学习器数量 N ,可以一定程度上提升 GvRF 算法的分类准确率,在较小基学习器数量时,快速增加模型的收敛速度,但在某些数据集中其收敛速度明显小于 XGBoost 等算法.

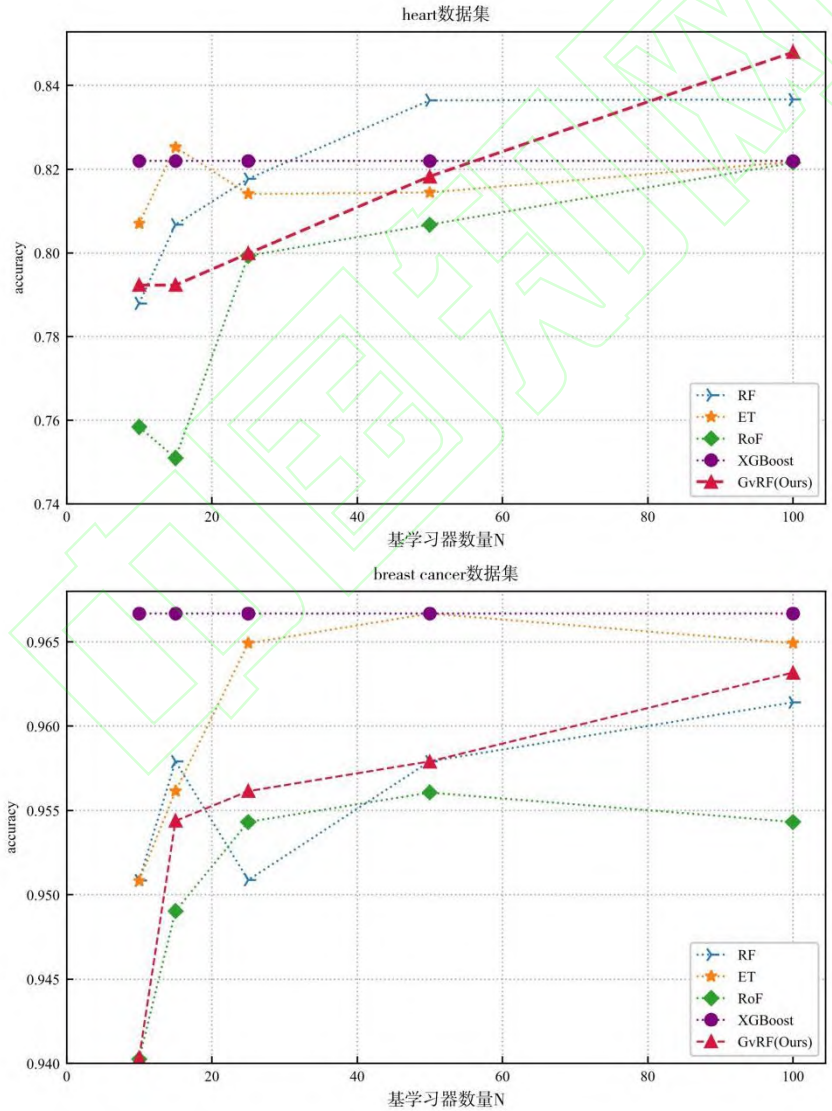

图 4 不同基学习器数量下的性能对比

Fig.4 Performance comparison of different N

4 结束语

本文通过分析随机森林算法,结合相似度粒化理论,在单特征上构造粒子,在多特征上由粒子形成粒向量,定义粒子与粒向量的大小、距离和运算方法.将相似度粒化技术引入随机森林算法中,设计基于粒向量的随机森林算法.由于粒子具有多角度、可并行的特点,通过随机选取参照样本构造多个相关系数较低的基学习

器,可以提高随机森林算法的性能.最后在多个不同类型数据集上进行实验,充分验证了文章所提出算法的正确性与有效性.未来阶段将重点针对以下几个问题展开研究:

(1)现阶段粒化理论存在一定的随机性,下一阶段将深入研究更具鲁棒性的粒化算法,构建包含更丰富的约束条件和先验知识的算法框架,采用非监督学习等方法,增强算法的泛化能力和鲁棒性.

(2)现阶段粒化理论计算成本相对较高,下一步将采用分布式计算等方法降低算法的计算复杂度,探索更合适的数据结构和搜索策略来优化算法的时间和空间复杂度.

(3)未来需要进一步提升提出的算法对基学习器的适应性,以拓宽其应用场景.后续研究将继续深入研究,提出算法的理论框架,丰富算法的理论基础.同时在更广泛的数据集和应用场景上验证算法效果,发现算法的潜在问题,不断改进算法,提高算法的精度、泛化能力,丰富算法的功能.

参考文献:

[1] ZADEH L A, GUPTA M M, RAGADE R K, et al. Fuzzy sets and information granularity[M]. 1979.

[2] LIN T Y. Neighborhood systems and relational data- bases[C]//Proceedings of the 1988 ACM sixteenth annual conference on Computer science. 1988: 725.

[3] LIN T Y, ZADEH L A. Special issue on granular computing and data mining[J]. International Journal of Intelligent Systems, 2004, 19(7): 565-566.

[4] YAO Y Y. Information granulation and rough set approximation[J]. International Journal of Intelligent Systems, 2001, 16(1): 87-104.

[5] YAO Y Y. Relational interpretations of neighborhood operators and rough set approximation operators[J]. Information sciences, 1998, 111(1-4): 239-259.

[6] 苗夺谦, 范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践, 2002, 22(1): 48-56.

MIAO D Q,FAN S D. The Calculation of Knowledge Granulation and Its application[J]. Systems Engineering - Theory & Practice. 2002, 22(1): 48-56.

[7] 胡清华,于达仁,谢宗霞.基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, 19(3): 640-649.

HU Q H, YU D R, XIE Z X. Numerical attribute reduction based on neighborhood granulation and rough approximation[J]. Journal of Software, 2008, 19(3): 640- 649.

[8] HU Q H, YU D R, XIE Z X Neighborhood classifiers[J]. Expert Systems with Applications, 2008, 34(2):866-876.

[9] ZHU P F, HU Q H, HAN Y H, et al. Combining neighborhood separable subspaces for classification via sparsity regularized optimization[J]. Information Sciences, 2016, s 370–371(1):270-287.

[10] CHEN Y, ZHU S, Li W, et al. Fuzzy granular convolutional classifiers[J]. Fuzzy Sets and Systems, 2022, 426: 145-162.

[11] LI W, MA X Y, CHEN Y M, et al. Random Fuzzy Granular Decision Tree[J]. Mathematical Problems in Engineering, 2021, 2021(10):1-17.

[12] CHEN Y, QIN N, LI W, et al. Granule structures, distances and measures in neighborhood systems[J]. Knowledge-Based Systems, 2019, 165: 268-281.

[13] 卜东波,白硕,李国杰.聚类/分类中的粒度原理[J]. 计算机学报, 2002, 25(8):810-816.

BU D B, BAI S, LI G J. Principle of granularity in clustering/classification[J]. Chinese Journal of Computers, 2002, 25(8):810-816.

[14] 王国胤,张清华,胡军.粒计算研究综述[J]. 智能系统学报, 2007, 2(6):8-26.

WANG G Y, ZHANG Q H, HU J. An overview of granular computing[J]. CAAI Transactions on Intelligent Systems, 2007, 2(6):8-26.

[15] DAI J, HAN H, ZHANG X, et al. Catoptrical rough set model on two universes using granule-based definition and its variable precision extensions[J]. Information Sciences, 2017(390):70-81.

[16] ROH S B, OH S K, PEDRYCZ W. A fuzzy ensemble of parallel polynomial neural networks with information granules formed by fuzzy clustering[J]. Knowledge-Based Systems, 2010, 23(3):202-219.

[17] WU W Z, LEUNG Y. Theory and applications of granular labelled partitions in multi-scale decision tables[J]. Information Sciences, 2011, 181(18):3878-3897.

[18] BREIMAN L. Random Forest[J]. Machine Learning, 2001, 45:5-32.

[19] 任家东,刘新倩,王倩,何海涛,赵小林.基于 KNN 离群点检测和随机森林的多层入侵检测方法[J].计算机研究与发展,2019,56(03):566-575.

REN J D, LIU X Q, WANG Q, HE H T, ZHAO X L. A multilevel intrusion detection method based on KNN outlier detection and random forests[J]. Journal of Computer Research and Development, 2019, 56(3): 566- 575.

[20] 沈晶磊,虞慧群,范贵生等.基于随机森林算法的推荐系统的设计与实现[J].计算机科学,2017,44(11):164-167.

SHEN J L, YU H Q, FAN G S, et al. Design and implementation of a recommendation system based on random forest algorithm[J]. Computer Science, 2017, 44(11): 164-167.

[21] 李旭明,李传军.基于随机森林模型的输电线路故障检测系统研究[J].计算技术与自动化,2020,39(01):29-33.

(Li X M, Li C J. Research on transmission line fault detection system based on random forest model[J]. Com- puting Technology and Automation, 2020,39(01):29-33.)

[22] 方匡南,吴见彬,朱建平,谢邦昌.随机森林方法研究综述[J].统计与信息论坛,2011,26(03):32-38.

FANG K N, WU J B, ZHU J P, XIE B C. A review of random forest method [J]. Statistics and Information Forum, 2011, 26 (03): 32-38.

[23] 徐计,王国胤,于洪.基于粒计算的大数据处理[J].计算机学报,2015,38(08):1497-1517.

XU J, WANG G Y, YU H. Big data processing based on granular computing[J]. Chinese Journal of Computers, 2015,38(08):1497-1517.

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

[25] RODRÍGUEZ D, JUAN J, LUDMILA I. K. and CARLOS J. A., Rotation Forest: A new classifier ensemble method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28: 1619-1630.

[26] 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. 2016: 785-794.