https://doi.org/10.1051/jnwpu/20224061261

一种基于最优集成随机森林的小样本数据特征提取方法

张维, 张浩晨

(西北工业大学 机电学院, 陕西 西安 710072)

摘 要:高维小样本数据作为数据挖掘的难点,用传统的随机森林算法进行特征选择时极易出现分类结果过拟合而导致的特征重要度排序稳定性差、精度低等问题。针对随机森林在小样本数据降维过程中出现的难点,提出了一种基于小样本数据特征提取算法 OTE-GWRFFS。基于生成对抗网络 GAN 进行样本扩充,避免传统随机森林在小样本分类过程中的过拟合现象;在数据扩充的基础上采用基于权重的最优树集合算法,减小生成数据分布误差对特征提取精度的影响,提升决策树集合的整体稳定性;采用单棵决策树的权重与特征重要性度量值加权平均得到特征重要性排序,从而解决了小样本数据特征选择过程中精度低稳定性差的问题。通过 UCI 数据集将所提算法与传统随机森林以及基于权重的随机森林算法进行实验对比,OTE-GWRFFS 算法在处理高维小样本数据时具有更高的稳定性和精度。

关 键 词:高维小样本数据;最优树集合;随机森林;特征提取;数据扩充

中图分类号:TP387 **文献标志码**:A **文章编号**:1000-2758(2022)06-1261-08

在特征选择算法的研究中, 随机森林(random forest,RF)算法[1]在构建决策树的过程中基于不纯度降低方法(mean decrease impurity)给特征进行排序。基于随机森林的 RVS 特征选择算法[2],利用给特征加入噪声后的袋外数据(out of bag data)预测准确率与原始袋外数据预测准确率之差,描述特征重要性度量值,因其能够有效去除冗余信息、处理时间短且效率高,成为特征选择的主流方法[3]。以上 2 种传统的随机森林特征选择法在处理数据量充足且分布均衡的数据时特征提取效果良好,但是在处理小样本数据时由于训练样本少,构建决策树分类时极易出现过拟合[4],导致计算特征重要性度量值时不稳定且精度低,降低了特征提取的有效性。

采用随机森林对高维小样本数据进行特征提取的难点主要在于样本量缺乏以及决策树的稳定性差、精度低。徐少成等[5]提出了基于随机森林的加权特征选择算法,采用单棵决策树的权重与特征重要性度量值加权平均得到最终的特征重要性度量值,提升了随机森林特征提取的精度,但是该方法在处理小样本数据时不能有效提升决策树集合的稳定性[6]。Khan 等[7]总结了最优树集合思想(optimal tree ensemble),认为随机森林中单株树的预测精度对最终分类有极大影响,提出了一种基于个体准确性和多样性的决策树选择方法,提升了随机森林分类的精度以及稳定性,该方法应用于小样本数据特征提取中可大大提升决策树集合的稳定性,但是该方法并未考虑小样本所带来的过拟合和特征提取精度低问题[8]。本文结合小样本数据在特征提取过程中出现的难点,提出了 OTE-GWRFFS 算法,结合生成式对抗网络(GAN)生成相似数据[9],并采用改进的非局部均值去噪算法[10]修正生成数据的分布误差,利用基于权重的最优树算法计算特征的重要性度量值,提高了小样本数据特征提取的精度、稳定性以及有效性。

收稿日期:2022-03-07

作者简介:张维(1970—),西北工业大学副教授,主要从事智能制造、制造数据分析技术研究。e-mail:zhangw@nwpu.edu.cn

1 OTE-GWRFFS 算法

随机森林在对高维小样本数据进行分类过程中,存在因样本量的缺乏导致训练深度不够以及过拟合现象。而利用袋外数据进行特征重要性度量值计算时,又有可信度低以及特征排序稳定性差问题。针对小样本数据在特征提取过程出现的问题,本文提出了一种基于最优集成随机森林的小样本数据特征提取方法(OTE-GWRFFS 算法)。OTE-GWRFFS 算法的流程如图 1 所示。

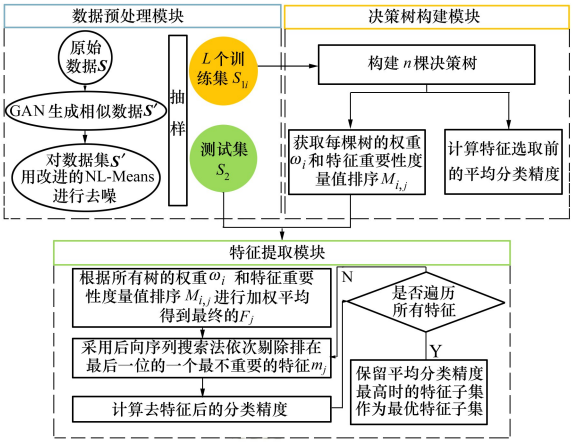

图 1 OTE-GWRFFS 算法流程图

1.1 OTE-GWRFFS 算法流程

OTE-GWRFFS 算法的具体步骤如下所示。

输入:原始数据集 $\boldsymbol{S}$;包括样本数 N 和样本特征 $\boldsymbol{A}=(A_1,A_2,A_3,\cdots,A_M)$

构建的随机森林的决策树个数 n

输出:样本特征 A_j 的最终特征重要性度量值 F_j

算法:

step1 初始化原始数据集 $\boldsymbol{S}$,设置构建随机森林的决策树个数 n

step2 利用 GAN 算法对原始数据集进行数据增强,得到生成数据集 $\boldsymbol{S}'$

step3 利用改进的 NL-Means 对生成数据集 $\boldsymbol{S}'$ 进行个别离群点的拟合得到数据集 $\boldsymbol{S}''$

step4 采用 bagging 抽样,构成 L 个训练数据集 $\boldsymbol{S}_{1i}(i=1,2,\cdots,L)$,一个测试数据集 $\boldsymbol{S}_2$,每个训练数据集中有 N' 个样本数据,m 个样本特征

step5 对数据集 $\boldsymbol{S}_{1i}$ 进行决策树的构建

step6 for (i = 1 to n)

计算所有决策树的精度 A_i

$$A_i=\frac{\sum_{s=1}^{S}t_{i,s}}{T} \tag{1}$$

式中:A_i 为第 i 棵决策树的分类精度;T 为测试集样本数量;$t_{i,s}$ 为第 i 棵决策树对测试样本的分类与样本真实分类相同的样本数目

将所有树的 A_i 按照由大到小排序

for(i = 1 To n)

逐次去除后 n'棵树并计算最终的分类精度 A

if (A 减小)

break

step7 for(i=1 to $n-n'$)

计算决策树的权重 ω_i

$$\omega_i=\frac{\sum_{e=1}^{E}t_{i,e}}{T} \tag{2}$$

式中:ω_i 代表第 i 棵树的权重;$t_{i,e}$ 代表第 i 棵决策树对测试样本的分类与随机森林所有树对测试样本的分类相同的样本数目;T 代表测试集的样本数量。

计算第 i 棵决策树中第 j 特征的重要性度量值 $M_{i,j}$

$$M_{i,j}=\frac{|A_i-A_{c,j}|}{\sum_{j=1}^{m}|A_i-A_{c,j}|} \tag{3}$$

式中,$A_{c,j}$ 定义为给测试集中第 j 个特征加入高斯噪声后的平均分类精度。

step8 最终的特征重要性度量值 F_j

$$F_j=\frac{\sum_{i=1}^{n}\omega_i\times M_{i,j}}{n-n'} \tag{4}$$

在计算完特征重要性度量值后,需要摒弃重要程度不高的特征,即采用后向搜索法,逐一去除重要程度靠后的特征并计算去除该特征后的平均分类精度,保留剩余最优特征子集的评判标准即去除该特征及排名低于该特征的其他特征后的平均分类精度达到最高。

1.2 时间复杂度分析

本文所提出的 OTE-GWRFFS 算法中基分类器选择 CART 算法。假设本算法中训练数据集的特征维数为 M,训练样本个数为 N,随机森林在构建 CART 树的过程中,从 N 个训练样本中利用 bagging 抽取 N' 个训练样本,从 M 个特征中随机选择 m 个特

征计算信息增益,并且对树的生长不进行剪枝。在本实验中,采用序列后向选择策略进行最优树的选择需要循环$(n-p)$次,特征选择需要循环$(m-p)$次(p由数据集特征数决定,一般不少于 5 个),根据排序后的特征集合生成新的训练数据集需要进行$(m-p)$次计算,每次计算时间为常数,故本算法总的时间复杂度可以近似表示为

$$O[(m-p)(n-p)\times O((m-p)N'(\lg N')^2)+m-p]\approx O[(n-p)(m-p)^2N'(\lg N')^2] \tag{5}$$

由(5)式可见,OTE-GWRFFS 算法的时间复杂度与特征维数 m 呈近似平方关系,与训练数据集样本个数 N'呈近似立方关系,对于高维小样本数据,运算时间是可以接受的,算法具有较好的扩展性。

2　小样本数据扩充

数据量小在特征选择过程中是影响特征排序精度以及稳定性的主要原因,从算法层面改进超参数设定的方法始终存在局限性[11],从数据层面通过数据增强技术扩充数据解决数据量小的问题,是提高特征选择精度以及稳定性的一种有效方法。本文依据表格数据与图像之间的等价性利用 GAN 对小样本数据集进行数据扩充。由于生成的数据在反归一化过程中会因小程度的偏差最终反映出较大的偏离,本文采用基于表格数据的非局部均值算法(NLM)对生成数据进行修正,提高生成数据与原始数据之间的分布相似性。数据生成模型如图 2 所示。

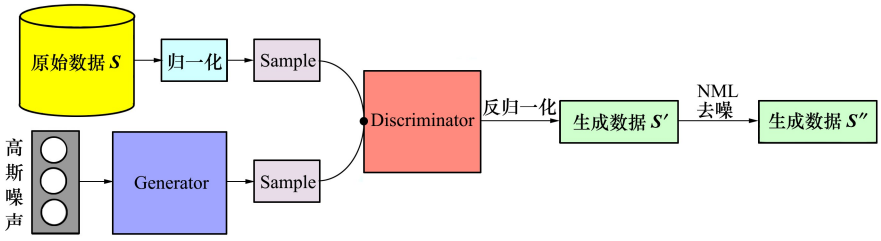

图 2　数据生成模型图

1) 数据生成

任意高斯噪声序列通过生成器将原低维向量投影成为与原始数据维度一致的高维向量,通过判别器将归一化后的原始数据 $\boldsymbol{S}$ 与生成的高维向量进行分布相似性判断,经过不断训练生成器和判别器,最终可以生成分布相似的数据,再经反归一化得到生成数据 $\boldsymbol{S}'$。

2) 数据分布修正

经过 GAN 网络生成的数据其分布与原始数据存在一定程度偏离,本生成模型通过改进后的 NLM 算法将生成数据与原始数据做相似性修正得到生成数据 $\boldsymbol{S}''$。最终待修正元素 X_j' 的最终值 Y_j 的计算公式如(6)式所示

$$Y_j=\sum_{i\in h}\psi_{ij}X_i \tag{6}$$

式中

$$\psi_{ij}=\frac{(1-D)^2}{2} \tag{7}$$

$$D=\sqrt{|X_j'-X_i|} \tag{8}$$

该数据生成模型通过提升生成数据 $\boldsymbol{S}'$ 和原始数据 $\boldsymbol{S}$ 的数据分布相似性解决了表格数据生成的偏离问题,同时数据量的扩增也避免了过拟合现象,提升了特征排序精度以及稳定性。

3　基于权重的最优树集合

在生成对抗网络的小样本数据扩充后,扩充样本与原始数据集有着一定的偏离性,不能真实地还原原始数据集的特征分布,因此在训练随机森林的过程中,某些决策树在随机分配训练数据集时被分到过多的生成数据,导致在测试数据集中分类效果极差[12],影响了特征重要性度量值的准确度,为了避免该问题的出现,本文采用集合高精度以及高多样性基分类器的方法,将训练好的基分类器按照分类精度排序并选取精度高的分类器作为最优决策树集合,可以在不影响决策树多样性的前提下降低不同类型模型的归纳偏差。

1) 分类错误率

为了挑选出分类性能最优的树,每棵树对测试集的分类错误率(分类精度)按(1)式计算。

根据所计算出每棵树的分类错误率(分类精度)将所有树的 A_i 按照由大到小排序,按照后向搜索法逐次选取前 n' 棵树并计算最终的平均分类精度 A,选取终止条件为最终平均分类精度 A 呈现下降趋势且基分类器数量有集成决策意义即可。

2) 特征重要性度量

在得到最优树集合后,对决策树给予不同权重再次综合评估特征重要性度量值。原始数据集 $\boldsymbol{S}$ 通过 bagging 抽样后会获得 L 个训练样本集 $\boldsymbol{S}_{1i}(i = 1, 2, \cdots, L)$ 和一个样本数为 T 的测试集 $\boldsymbol{S}_2$,这 n 个训练样本集可以产生 n 棵决策树,根据决策树的预测结果可以获得一个 $T \times (n + 2)$ 的矩阵,如表 1 所示。

表 1 $n=7, T=5$ 的决策矩阵

序号	TR1	TR2	TR3	TR4	TR5	TR6	TR7	决策结果	原始分类
1	1	0	1	1	1	1	0	1	1
2	0	1	0	0	1	1	0	0	0
3	1	1	0	1	0	1	1	1	1
4	1	0	1	1	0	0	1	1	1
5	0	0	0	1	1	0	1	0	1

表 1 中第 i 棵决策树的可信度(权重)可由(9)式得到

$$\omega_i = \frac{\sum_{e=1}^{E} t_{i,e}}{T} \times A_{\text{ensemble}} \tag{9}$$

式中:$t_{i,e}$ 代表第 i 棵树中对 T 个测试样本的分类结果和决策结果中一致的数量,A_{ensemble} 表示集成预测的准确率,即决策结果与原始分类的相符程度。由于每棵决策树的 A_{ensemble} 值都是一样的,是否考虑 A_{ensemble} 的作用对排序结果没有影响,在计算权重时加入这个因素,其目的是尽量缩小各决策树间权重的相对差距。

4 实验验证

为验证本 OTE-GWRFFS 算法在高维小样本数据集上的有效性,在 UCI 数据集中挑选了 5 个具有代表性的小样本数据集。对于每个数据集,都首先利用 GAN 进行样本扩充,样本扩充的原则即保证原始数据集的分布特征,样本扩充量不能超过原始数据集的样本数,保证了在 bagging 抽样时训练集有足够充分的原始样本。表 2 列出了这些数据集名称、特征以及数据扩充结果。

表 2 UCI 中小样本实验数据汇总

序号	数据集	样本个数	特征数	分类数目	样本扩充量	扩充后样本个数
1	dataR2	116	9	2	84	200
2	glass	214	9	7	86	300
3	parkinsons	195	22	2	105	300
4	wpbc	194	33	2	106	300
5	sonar	208	60	2	92	300

为了验证最优树集合算法(OTE)的有效性,图 3 展示了 5 个数据集在后向搜索法过程中摒弃分类精度低的决策树后对测试集的分类精度。可以看出每个数据集在选择最优树过程中都有一个精度峰值,此峰值所对应的决策树量即最优树数量,表明最优树集合算法(OTE)可以有效选择出分类精度最高的树。

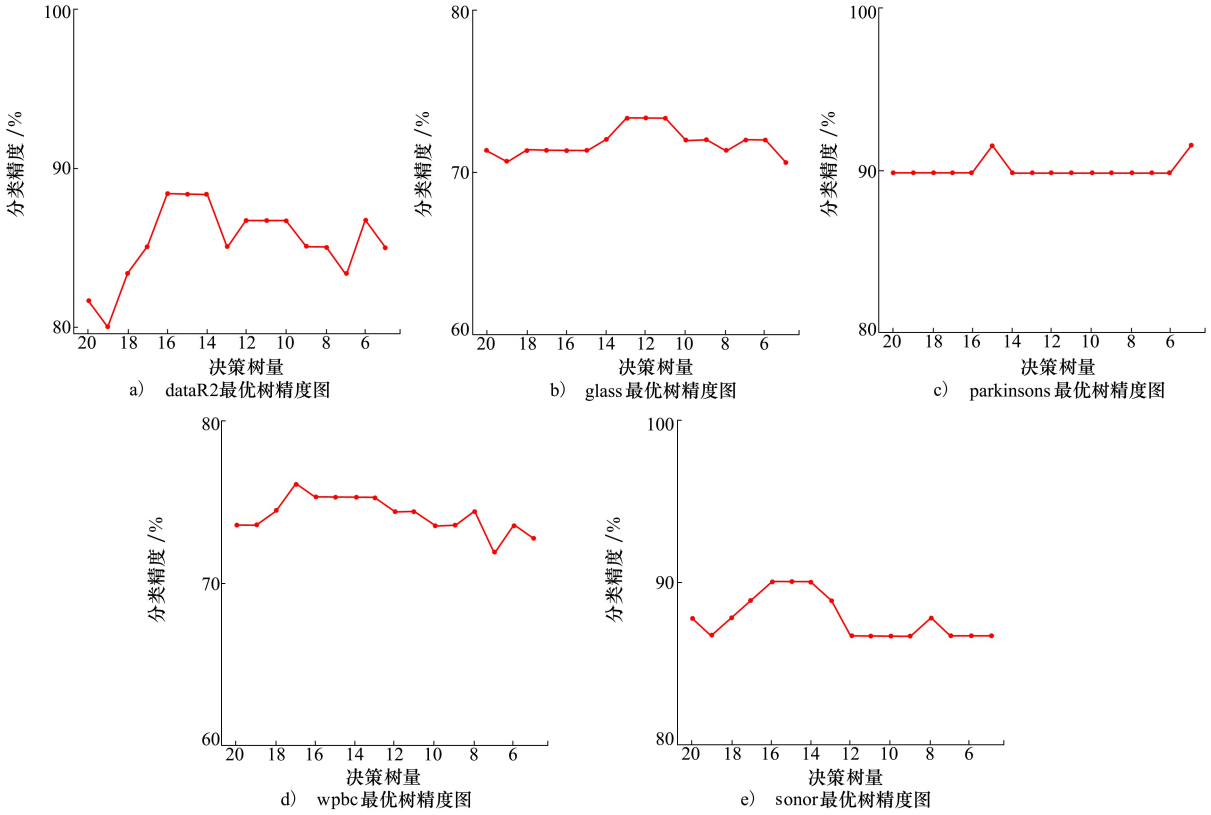

图 3　最优树集分类精度图

表 3 列出了所用数据集在利用 RFFS、WRFFS 以及 OTE-GWRFFS 算法进行特征选择后得到的特征子集个数,选择依据为各算法对特征重要性度量值进行排序并采用后向搜索法选择后分类精度达到最高时的特征子集个数。经过特征子集个数的筛选,RFFS、WRFFS 以及 OTE-GWRFFS 算法的维数平均下降率分别为 25.68%,25.04%和 34.20%。

表 3　各算法特征选择结果

数据集	特征数	各算法得到的特征子集个数		
		RFFS	WRFFS	OTE-GWRFFS
dataR2	9	7	6	6
glass	9	7	7	7
parkinsons	22	12	13	10
wpbc	33	28	27	25
sonar	60	36	37	28

表 4 给出了所有数据集在进行特征选择前的分类精度,以及利用 RFFS、WRFFS 和 OTE-GWRFFS 算法进行特征选择后,再次对特征选择后的数据集进行分类,经过对比,RFFS、WRFFS 以及 OTE-GWRFFS 算法的分类精度平均提升率分别为 7.91%,9.42%和 13.39%。

表 4　各算法特征选择前后分类精度对比　%

数据集	特征选择前分类精度	各算法特征选择后分类精度		
		RFFS	WRFFS	OTE-WRFFS
dataR2	52.17	55.88	58.82	61.76
glass	67.69	71.88	71.88	75.00
parkinsons	91.40	93.10	94.83	94.83
wpbc	72.88	83.05	83.05	86.44
sonar	80.48	88.89	88.89	90.48

表 5 给出了所有数据集在未进行特征提取前以及经过各算法特征提取后的 F1-score 值。

为了清楚表达本算法在特征提取方面优于 RFFS、WRFFS 算法,图 4 展示了 3 种算法在特征提取过程中随着特征数的降低,其分类精度的变化曲线。

表 5 各算法特征选择前后的 F1-score 值

数据集	特征选择前	特征选择后 F1-score		
	F1-score	RFFS	WRFFS	OTE-GWRFFS
dataR2	0.507 9	0.565 8	0.627 4	0.732 5
glass	0.539 9	0.549 0	0.549 0	0.598 9
parkinsons	0.873 0	0.873 0	0.873 0	0.873 0
wpbc	0.410 0	0.415 8	0.484 6	0.475 6
sonar	0.775 0	0.744 4	0.840 3	0.856 6

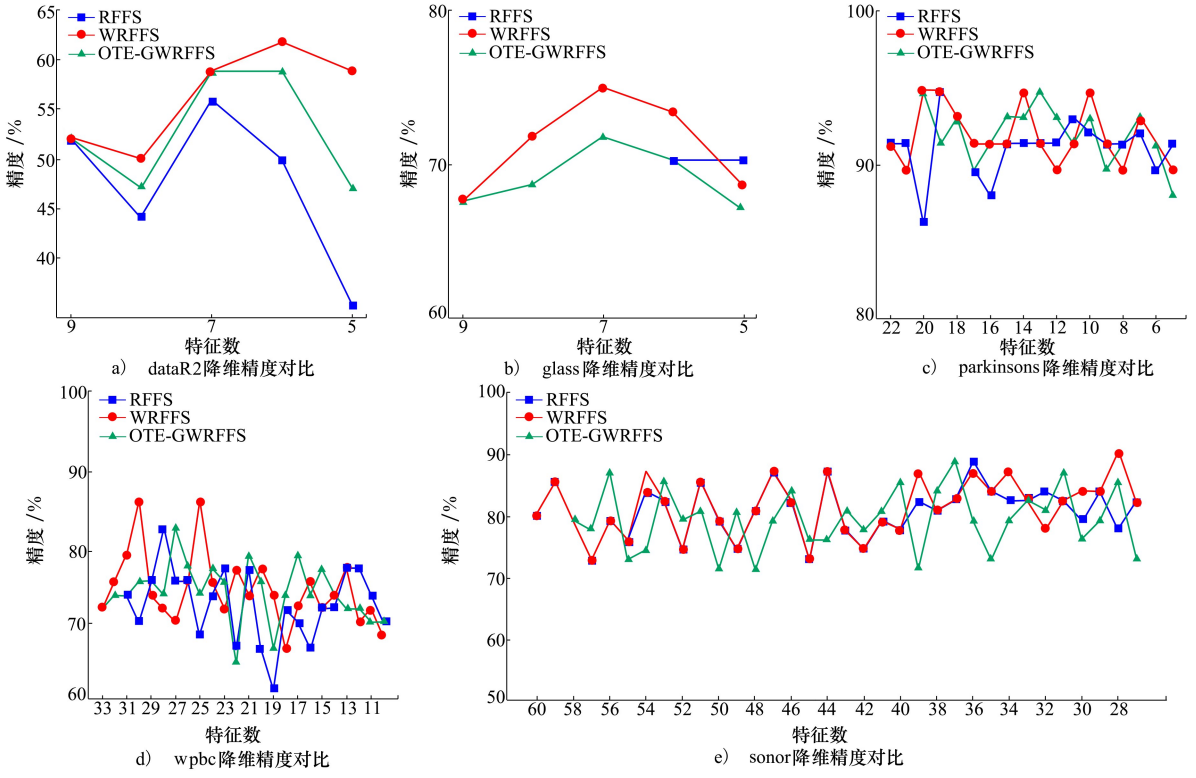

图 4 降维精度对比图

表 6 给出了数据扩充算法前后的最优特征子集数和分类精度对比,经数据扩充后算法的维数平均筛除提升率为 9.16%,分类精度平均提升率为 6.96%,验证最优树集合算法对小样本数据扩充的有效性。

表 6 数据扩充前后 OTE 算法的特征选择精度对比

数据集	特征选择前分类精度/%	基于 OTE 思想			
		未扩充的特征选择		扩充后的特征选择	
		特征数	分类精度/%	特征数	分类精度/%
dataR2	52.17	6	61.76	6	61.76
glass	67.69	8	70.31	7	75.00
parkinsons	91.40	15	96.55	10	94.83
wpbc	72.88	24	74.41	25	86.44
sonar	80.48	47	82.25	28	90.48

根据对表 3 的特征选择结果和表 4 的特征选择前后分类精度对比以及表 5 计算的分类 F1-score 值可知,本算法在基于相同的数据集进行特征选择后维数的平均下降率为 34.20%,而 RFFS 以及 WRFFS 算法的维数平均下降率大约仅有 25%,且经过本算法特征降维后再次分类的 F1-score 值达到最高。可以证明本文算法相比于 RFFS 以及 WRFFS 算法有较大提升。表 4 展示了删除冗余特征后本算法再次进行分类,分类精度提升率达到 13.39%,而 RFFS 以及 WRFFS 算法的分类精度提升率大约仅有 8%,同时在特征提取后进行再次分类的 F1-score 值有明显提升,说明本算法能够最大程度地对特征进行降维处理,能够更有效地删除冗余特征,并且特征选择精度更高。表 6 用数据扩充前后的分类精度作为对比,可以看出在用于验证的数据集中,数据扩充对维数平均筛除提升率约为 9.16%,分类精度的提升率大约在 6.96%,可以证明数据扩充在处理小样本数据时有效地提升了特征选择的精度以及稳定性。

结合实验结果,在特征选择的维数平均下降率以及在分类精度方面,本算法都比其他 2 种算法更有效、精度更高。由于选取的数据具有广泛的代表性,所以说本算法在特征选择上具有更强的适用性。且本算法在针对于极小样本数据集时也具有有效性,可以完全避免过拟合现象且特征提取效果良好。

5　结　论

高维小样本数据的特征降维极容易出现特征排序不稳定,经常会将关键特征作为不重要特征处理,大大影响了降维的精度,不利于后续数据挖掘工作。本文基于小样本的降维问题,提出了基于最优集成随机森林的小样本数据特征提取方法 OTE-GWRFFS。建立数据增强模型,在数据扩充的基础上,采用改进的 NL-Means 去噪法以及最优树集合 OTE 思想改善数据扩增过程中出现的数据偏差性质,通过给予最优树集合以不同权重再次提升每棵决策树的重要性度量的可靠性。实验表明 OTE-GWRFFS 算法可以避免随机森林过拟合问题,提升了特征排序的稳定性及精度,在经过特征选择后,随机森林分类精度明显提升。

参考文献:

[1] HASSAN H, BADR A, ABDELHALIM M B. Prediction of o-glycosylation sites using random forest and GA-tuned PSO technique[J]. Bioinformatics & Biology Insights, 2015, 9(9): 103-109

[2] ROBIN G, JEAN-MICHEL P, CHRISTINE T. Variable selection using random forests[J]. Pattern Recognit, Lett, 2010, 31: 2225-2236

[3] 姚登举, 杨静, 詹晓娟. 基于随机森林的特征选择算法[J]. 吉林大学学报, 2014, 44(1): 137-141
YAO Dengju, YANG Jing, ZHAN Xiaojuan. Feature selection algorithm based on random forest[J]. Journal of Jilin University, 2014, 44(1): 137-141 (in Chinese)

[4] 王翔, 胡学钢. 高维小样本分类问题中特征选择研究综述[J]. 计算机应用, 2017, 37(9): 2433-2438
WANG Xiang, HU Xuegang. A review of feature selection in high-dimensional small sample classification[J]. Computer Application, 2017, 37(9): 2433-2438 (in Chinese)

[5] 徐少成, 李东喜. 基于随机森林的加权特征选择算法[J]. 统计与决策, 2018, 34(18): 25-28
XU Shaocheng, LI Dongxi. Weighted feature selection algorithm based on random forest[J]. Statistics and Decision Making, 2018, 34(18): 25-28 (in Chinese)

[6] LI H B, WANG W, DING H W, et al. Trees weighting random forest method for classifying high dimensional noisy data[C]// IEEE 7th International Conference on E-Business Engineering, 2010

[7] KHAN Zardad, ASMA Gul, ARIS Perperoglou, et al. Ensemble of optimal trees, random forest and random projection ensemble classification[J]. Advances in Data Analysis and Classification, 2020, 14: 97-116

[8] KHAN Z, GUL A, MAHMOUD O, et al. An ensemble of optimal trees for class membership probability estimation// Analysis of large and complex data[M]. Switzerand: Springer International Publshiug, 2016: 395-409

[9] WEN B, LUIS O, COLON K P. Subbalakshmi and ramamurti chandramouli causal-TGAN: generating tabular data using causal generative adversarial networks[D]. Hoboken: Stevens Institute of Technology, 2021

[10] 赵庆平, 陈得宝, 姜恩华, 等. 一种改进权重的非局部均值图像去噪算法[J]. 电子测量与仪器学报, 2014, 28(3):

334-339
ZHAO Qingping, CHEN Debao, JIANG Enhua, et al. An improved weighted nonlocal mean image denoising algorithm[J]. Journal of Electronic Measurement and Instrument, 2014, 28(3): 334-339 (in Chinese)
[11] KUNCHEVA L I, MATTHEWS C E, ARNAIZ-GONZÁLEZ A, et al. Feature selection from high-dimensional data with very low sample size: a cautionary tale[J/OL]. (2020-08-27)[2022-01-19]. https://arxiv.org/abs/2008.12025
[12] 李秋玮. 基于条件生成对抗网络和超限学习机的小样本数据处理方法研究[D]. 镇江: 江苏大学, 2019
LI Qiuwei. Research on small sample data processing method based on conditional generation countermeasure network and transfinite learning machine[D]. Zhenjiang: Jiangsu University, 2019 (in Chinese)

A feature extraction method for small sample data based on optimal ensemble random forest

ZHANG Wei, ZHANG Haochen
(School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: High dimensional small sample data is the difficulty of data mining. When using the traditional random forest algorithm for feature selection, it is to have the poor stability and low accuracy of feature importance ranking caused by over fitting of classification results. Aiming at the difficulties of random forest in the dimensionality reduction of small sample data, a feature extraction algorithm ote-gwrffs is proposed based on small sample data. Firstly, the algorithm expands the samples based on the generated countermeasure network Gan to avoid the over fitting phenomenon of traditional random forest in the small sample classification. Then, on the basis of data expansion, the optimal tree set algorithm based on weight is adopted to reduce the impact of data distribution error on feature extraction accuracy and improve the overall stability of decision tree set. Finally, the weighted average of the weight and feature importance measure of a single decision tree is used to obtain the feature importance ranking, which solves the problem of low accuracy and poor stability in the feature selection process of small sample data. Through the UCI data set, the present algorithm is compared with the traditional random forest algorithm and the weight based random forest algorithm. The ote-gwrffs algorithm has higher stability and accuracy for processing high-dimensional and small sample data.

Keywords: high dimensional small sample data; ensemble of optimal trees; random forest; feature extraction; data expansion

引用格式: *张维, 张浩晨. 一种基于最优集成随机森林的小样本数据特征提取方法[J]. 西北工业大学学报, 2022, 40(6): 1261-1268*
ZHANG Wei, ZHANG Haochen. A feature extraction method for small sample data based on optimal ensemble random forest[J]. Journal of Northwestern Polytechnical University, 2022, 40(6): 1261-1268 (in Chinese)

© 2022 Journal of Northwestern Polytechnical University.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.