基于适应性函数的自适应性遗传算法及其实验验证

李 雷[1,3], 苏明昕[2], 苏 波[2], 刘 刚[2]

(1.国电南瑞科技股份有限公司, 江苏 南京 211106;

2.国网宁夏电力有限公司, 宁夏 银川 750000;

3.国电南瑞南京控制系统有限公司, 江苏 南京 211106)

摘 要: 为了提高数据挖掘中从高维特征空间选择最优特征子集预测效果,构建得到了一种改进后的遗传算法来完成特征选择过程,并加入了多模型融合的分析方法,采用多点位交叉算法研究了不同特征组合产生的预测结果。测试结果显示此算法能够筛选得到和目标变量具有最大相关度以及可以高效区分目标值的特征。当特征维数减少后,预测算法所需的运行时间也显著降低。

关键词: 数据挖掘;特征选择;遗传算法;预测模型

中图分类号:TP184 文献标识码:A 文章编号:1003−7241(2020)08−0011−04

Adaptive Genetic Algorithm and It's Experimental Verification Based on Adaptive Function

LI Lei[1, 3], SU Ming-xin[2], SU Bo[2], LIU Gang[2]

(1. NARI Technology Co., Ltd., Nanjing 211106 China;

2. State Grid Ningxia Electric Power Co., Ltd., Yinchuan 750000 China;

3. NARI Nanjing Control System Co., Ltd., Nanjing 211106 China)

Abstract: In order to improve the prediction effect of selecting the optimal feature subset from the high-dimensional feature space in data mining, an improved genetic algorithm is constructed to complete the feature selection process, and the multi-model fusion analysis method is added. The prediction results of different feature combinations are studied using the multi-point crossover algorithm. The test results show that this algorithm can filter out the features that have the maximum correlation with the target variables and can effectively distinguish the target values. When the feature dimension is reduced, the running time of the prediction algorithm is also reduced significantly.

Key words: data mining; feature selection; Genetic Algorithm; prediction model

1 引言

进入信息化时代后,各类信息数据呈现海量增加的趋势。不同数据之间的联系也变得更加紧密。因此人们在建立机器学习模型时需要综合考虑不同数据特征的组合与交互过程。同时。随着各项业务的不断发展。众多个性化挖掘挖掘方法被应用于数据分析领域,从而使数据特征表现出明显的多样化特征[1−3]。在上述发展趋势下逐渐形成了非常庞大的机器学习模型特征空间。此时需要从大量的特征提取得到与目标变量可以测试很高关联度的特征,并尽量删除没有关联性的冗余特征。从而构建得到由许多子特征组成的最优空间[4−6]。对模型进行筛选处理可以提高其可释性,同时因为减小了特征空间的规模来达到降维的目的,可以利用该方法获得更快的训练速度并获得更高的预测精度。各学习模型所采用的特征选择方法也存在较大差异,根据不同的选择方式可将其分成包裹法、过滤法与嵌入法共三种[7]。以上各方法依次从特征组合方式、单一特征以及模型泛化能力角度实

收稿日期: 2018-11-28

现对各类不同特征的过滤[8]。还可把特征选择理解为求解原始特征空间的最优解，目前已有较多研究人员在特征选择方面引入了启发式搜索算法。由此实现对上述3种特征选择方法的综合应用，使各种方法的优势得以相互补充。采用上述各类遗传算法来分析种群适应值时，都是建立在单模型基础上。通常这种方法只能获得非常单一的适应函数，不能满足不同环境的使用要求[9−10]。根据以上分析，本文构建得到了一种改进后的遗传算法来完成特征选择过程，并加入了多模型融合的分析方法，采用多点位交叉算法研究了不同特征组合产生的预测结果。最后对本文遗传算法筛选特征子集的过程进行验证，分别测试了不同的UCI标准数据集与遗传次数下的结果。

2 基于特征选择的遗传算法

2.1 种群个体的编码

采用遗传算法把原始问题包含的各个可行解都模拟成相应的种群个体，算法的搜索空间范围取决于种群的规模。对种群实施初始化处理时，如果选择的种群规模偏大或偏小时则无法利用此算法来达到全局最优。在这种情况下可以设定种群M值介于30～80的合适范围内，再根据问题类型和难易性进行调节。

通过种群个体来表示各种可能的特征组合结果，以二进制的形式完成个体基因的编码。同时0与1来表示各特征的选中状态。利用不同的二进制串来表示个体基因，并且在Ni等于1的时候代表此特征被选中。反之Ni等于0。

本文在编码种群个体的过程中加入了而方差阈值的方法。把各特征的方差值和阈值实施比较，按照与阈值进行大小比较所得的结果来实现过滤原始特征空间的过程，这样便可以实现以统计学的方式将具有较小变化幅度与不同区分度的特征全部去除。通常情况下，此类特征对于模型的目标变量分类几乎不会发挥作用，利用最大方差筛选后能够显著降低种群个体基因串的规模，从而使遗传算法只需对更小范围进行搜索。最终算法迭代速度显著提高。

2.2 适应性函数

适应性函数属于遗传算法的重要组成部分，可以采用该函数来分析种群进化阶段面临的环境难易情况，考虑到遗传算法只需利用适应性函数便可以对种群质量进行有效调节。由此可见，合理选择适应性函数可以促使算法更快达到全局最优点的收敛状态。在特征选择的过程中，本文以预测模型的输出结果作为个体适应值。获得各特征组合条件下得到的预测拟合度。结果显示，随着适应值的增大，可以得到更优良的个体。同时，本研究还引入多折交叉的方式对结果进行了验证，由此减小个体对目标的过拟合程度。使个体适应性达到更高鲁棒性；综合运用融合线性分析模型以及树模型可以使特征组合实现更高的鲁棒性。根据以上分析可以将个体适应性值F(Xi)表示成如下形式：

$$F(X_i)=F(x_1,\cdots,x_n)=\frac{\sum_{j=1}^{cv}L(x_1,\cdots,x_n)}{cv}+\frac{\sum_{j=1}^{cv}G(x_1,\cdots,x_n)}{cv},\quad i=1,\cdots,M,n=1,\cdots,N \tag{1}$$

2.3 算法流程

根据特征选择构建的遗传算法是利用特征工程来完成特征空间的搜索。在处理初始数据时，本文采用热编码的方法来分析类型特征以及采用交叉组合模式分析数值特征。经过上述各方法处理后可以在未加入先验知识的条件下，使原始特征空间获得充分扩展，之后利用方差过滤为特征空间构建特征候选集，再对种群遗传进化过程进行初始化。当获得最大迭代次数时，再对最优特征子集进行反馈，利用上述子集完成预测分析，同时给出评价结果。从图1中可以看到上述方法的具体实现流程。

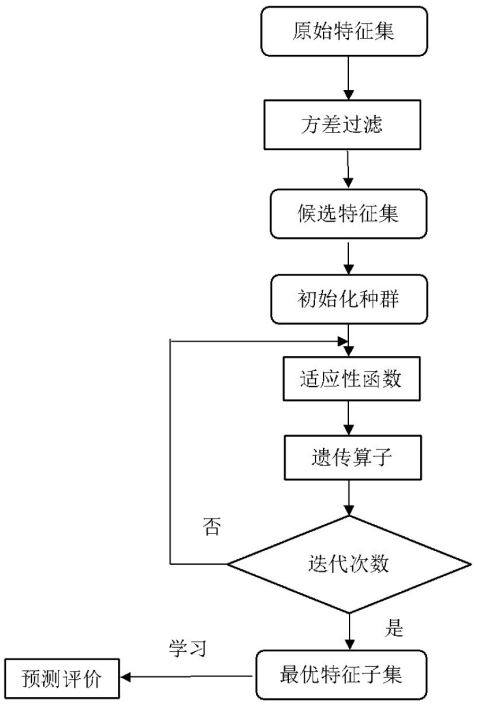

图1 遗传算法流程图

3 实验

3.1 实验环境

本文选择Python3.6.1版本来编写遗传算法的框架；通过相应的机器学习库来调用预测模型并处理各项数据特征；通过Ipython−notebook调试算法开发得到编译器。对UCI公开的四种标准测试数据集进行了测试，同时为了更加便于进行对比。把上述各数据集以d1～d4的形式分别进行表示，从表1中可以看到所有数据集包含的各指标对应的维数结果；本研究利用这些数据集测试得到遗传算法的输入特征，并构建了线性回归模型来完成预测分析。

表1 数据集特征信息

数据集	原始维度	工程维度	样本数
Facebook(d1)	20	48	510
Forestfires(d2)	15	58	526
Energy(d3)	32	75	19 750
Student(d4)	35	59	651

本文综合使用均方误差(MSE)和交叉验证的方法对遗传算法的各项特征预测结果进行了评价，得到如下所示的均方误差公式：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}-y)^2$$

上式的$\hat{y}$代表模型预测所得的结果。y代表样本真实值。

种群数量总共等于20。所有个体所拥有的基因长度都通过最大方差进行筛选得到。种群所能达到的最大遗传次数依次等于10次、20次与30次。将其表示为L1、L2、L3；设定最大方差阈值等于0.1。以默认参数完成适应性函数的线性回归和GBDT模型分析。总共进行10次交叉验证，之后利用线性回归(LR)方法来进行预测，对比了不同模型所得到的预测结，以较低的MSE值作为最佳方法。

3.2 实验结果

表2 方差筛选

数据集	d1	d2	d3	d4
特征数量	43	46	75	45

根据表2可知。采用最大方差方法来筛选上述数据集能够充分去除低区分度特征。测试结果显示，d1以及d3降低2个维度。表明这类数据集具有丰富的原始特征取值，可以实现对目标变量的精确区分与识别；其余数据集减少了约一半的维度，去除了具有单一取值的数据，使特征空间维度明显减小，实现搜索范围的明显压缩。

表3的Q代表对4个数据集经过L3遗传次数选择生成的剩余特征，相比之前特征空间发生了显著缩减。其中，d3数据集减少近50%，遗传算法可以得到更高准确率。M1和M2分别表示特征选择前后最终线性回归模型的MSE值，并且d1数据集具有最高的准确率。由此可见此算法能够筛选得到和目标变量具有最大相关度以及可以高效区分目标值的特征。

表3 筛选后特征数量与MSE

数据集	d1	d2	d3	d4
Q	24	23	36	25
M1	94.71	1.28	210.61	1.41
M2	981.61	1.53	211.54	1.48

表4显示了采用L3遗传次数进行特征选择前后得到的预测模型需要的运行时间。分别以T1和T2表示特征选择前与选择后的时间。可以明显发现，当特征维数减少后，预测算法所需的运行时间也显著降低。这是由于此时模型复杂度发生了下降，因此只需对各特征进行更少次数的遍历，最终使得此遗传算法可以实现预测准确率提升的前提下使预测效率也获得明显优化。

表4 各数据集运行效率

数据集	d1	d2	d3	d4
T1	0.025	0.028	1.682	0.268
T2	0.022	0.023	1.028	0.022

4 结束语

1）采用遗传算法把原始问题包含的各个可行解都模拟成相应的种群个体，通过种群个体来表示各种可能的特征组合结果，在编码种群个体的过程中加入了而方差阈值的方法。采用该函数来分析种群进化阶段面临的环境难易情况，根据特征选择构建的遗传算法是利用特征工程来完成特征空间的搜索。

2）对四种标准测试数据集进行了测试，测试结果显示d1以及d3降低2个维度。表3的Q代表对4个数据集经过L3遗传次数选择生成的剩余特征，得到d3数据集减少近50%；此算法能够筛选得到和目标变量具有最大相关度以及可以高效区分目标值的特征。当特征维数减少后，预测算法所需的运行时间也显著降低。

（下转第19页）

5 结束语

本文主要论述了力传感器的动态校准的方法与装置。首先介绍了力传感器在两种常规动态校准条件下力传感器的数学模型，并建立了对应的运动方程，说明了力传感器动态校准结果差异性以及产生差异的原因；然后介绍了一种基于刚体碰撞的高精度的脉冲力发生装置，说明了其结构和工作原理；同时也介绍了采用外插式激光干涉仪测量的方法与解算过程；最后介绍了典型力传感器在本装置上脉冲力校准的结果并对结果进行了初步分析。力传感器的动态校准及可靠量值溯源性，目前仍然是力学计量校准中动态计量研究的主要内容之一，包括基于力传感器模型参数辨识的校准、阶跃力校准力传感器、力传感器动态校准结果的评估等[13]，同时这也以此装置为基础进一步的内容。

参考文献：

[1] 张智敏，张跃，周宏，等.全自动1MN静重式力标准机[J].计量学报，2008，29(1)：65−68.

[2] 张智敏.大力值国际关键比对的评述[J].计量技术，2007(9)：47−50.

[3] 尹肖，张力，王宇，等.动态力校准中力传感器端部等效质量测量方法研究[J].计测技术，2014，34(6)：50−54.

[4] YUSAKU FUJII.Measurement of steep impulse response of a force transducer[J].Meas.Sci.Technol，2003(14)：65−69.

[5] ROLF KUMME.Investigation of the comparison method for the dynamic calibration of force transducer[J].Measurement，1998(23)：239−245.

[6] 商佳尚，王宇.动态力校准中需要规范的若干问题[J].计测技术，2014，4(28)：1−5，10.

[7] 国家质量监督检验检疫总局.JJF1370−2012正弦法力传感器动态特性校准规范[S].2012.

[8] 孟峰，张跃，张智敏，等.脉冲式动态力校准装置发展动态[J].计量技术，2011(5)：47−50.

[9] 商一奇，李善明，曹亦庆，等.脉冲力校准装置提升及扶正机构设计及数据研制[J].计测技术，2017，37(6)：35−39.

[10] CH SCHLEGEL，G KIECKENAP，B GLOCKNER，et al.Traceable periodic force calibration[J].Metrologia，2012(49)：224−235.

[11] 胡红波，孙桥.基于绝对法冲击校准的加速度计参数辨识的研究[J].测试技术学报，2013，27(1)：19−25.

[12] NICHOLAS VLAJIC，AKO CHIJIOKE.Traceable dynamic calibration of force transducers by primary means[J].Metrologia，2016(53)：136−148.

[13] C.BARTOLI，M.F.BEUG，T.S.BRUNS et al.Traceable dynamic measurement of mechanical quantities：objectives and first results of this European project[C].Proceedings of XX IMEKO World Congress，2012.

[14] MICHAEL KOBUSCH.Analysis of shock force measurements for the model−based dynamic calibration[Z].8thWorkshop on Analysis of Dynamic Measurements，2014.

[15] 国家质量监督检验检疫总局.GB/T 14412−2005机械振动与冲击加速度计的机械安装[S].2005.

作者简介：卢小犇(1985−)，男，工程师，研究方向：力学计量，振动冲击，非标在线校准。

(上接第13页)

参考文献：

[1] 桑田，张书海，胡啸.基于遗传算法改进的站点分布位置优化模型[J].中国战略新兴产业，2018(16)：1−2.

[2] 刘畅，谢文俊，张鹏，郭庆.多目标群多基地多无人机协同任务规划[J].弹箭与制导学报，2018(21)：1−6.

[3] 赵玉强，钱谦.基于最优加权进化的自适应遗传算法及性能研究[J].信息技术，2018(11)：68−72.

[4] 保玉俊，周莉莉，段鹏.一种基于遗传算法的加权朴素贝叶斯分类算法[J].云南民族大学学报(自然科学版)，2018(6)：525−529.

[5] 王佐勋，李亚洲，李国庆.基于遗传算法的LCL型逆变器双闭环参数设计[J].电力系统保护与控制，2018(161)：1−7.

[6] 司华清，袁亚琼.一种结合自适应惯性权重的改进遗传粒子群算法[J].技术与市场，2018，25(11)：48−49.

[7] 王艳，程丽军.基于事件驱动的云端动态任务分解模式优化方法[J].系统仿真学报，2018，30(11)：4029−4042.

[8] 林东升.一种基于遗传算法的小波阈值去噪方法[J].科技与创新，2018(21)：115−117.

[9] 郭慧，刘忠宝，柳欣.基于云模型和决策树的入侵检测方法[J].计算机工程，2018，30(11)：1−9.

[10] 朱卓悦，徐志刚，沈卫东，杨得玉.基于遗传蝙蝠算法的选择性拆卸序列规划[J].浙江大学学报(工学版)，2018(11)：1−8.

作者简介：李雷(1978−)，男，硕士，高级工程师，研究方向：智能电网调度自动化，新能源调度运行分析。