一种融合蚁群算法和随机森林的特征选择方法

李光华[1] 李俊清[1,2] 张 亮[1] 辛衍森[1] 邓华伟[1]
(山东农业大学信息科学与工程学院 山东 泰安 271018)[1]
(山东农业大学农业大数据研究中心 山东 泰安 271018)[2]

摘 要 面对海量的高维数据,剔除冗余特征来进行特征筛选,已成为当今信息与科学技术面临的重要问题之一。传统的特征选择方法不适合对整个特征空间进行搜索,其运行性能以及准确性低下。文中提出了一种融合蚁群算法和随机森林的特征选择方法,该方法将随机森林的重要度评分作为蚁群算法的启发式信息,然后采用蚁群算法进行智能搜索,并将特征选择结果作为评价指标实时反馈给蚁群的信息素。实验表明,该特征选择方法与传统特征选择方法相比,能够有效地减少数据集中的特征数量,同时提高了数据分类的准确率。

关键词 特征选择,蚁群算法,随机森林
中图法分类号 TP391 **文献标识码** A

Feature Selection Method Based on Ant Colony Optimization and Random Forest

LI Guang-hua[1] LI Jun-qing[1,2] ZHANG Liang[1] XIN Yan-sen[1] DENG Hua-wei[1]
(School of Information Science and Engineering, Shandong Agricultural University, Tai'an, Shandong 271018, China)[1]
(Agricultural Big Data Research Center, Shandong Agricultural University, Tai'an, Shandong 271018, China)[2]

Abstract In the face of massive high-dimensional data, eliminating redundant features for feature selection has become one of the important issues faced by information and science and technology today. Traditional feature selection methods are not suitable for searching the whole feature space, and their performance and accuracy are low. In this paper, a method of feature selection based on ant colony optimization and random forest was proposed. This method takes the importance score of random forest as the heuristic factor of ant colony optimization, uses ant colony optimization to search intelligently, and uses the result of feature selection as the evaluation index to feedback the pheromone of ant colony in real time. Experiments show that this feature selection method can effectively reduce the number of features in data sets and improve the accuracy of data classification compared with traditional feature selection methods.

Keywords Feature selection, Ant colony optimization, Random forest

1 引言

大数据时代,数据呈现高维性、非线性等复杂特征。现阶段机器学习模型的特征空间往往庞大且复杂,面对海量的高维数据,剔除冗余特征进行特征筛选,已成为当今信息与科学技术面临的重要问题之一。对于高维数据来说,在整个数据维度上往往很难找到反映数据特点的特征区域,多数传统的特征选择方法无法有效地对整个特征空间进行搜索,运行性能以及准确性较低。针对上述问题,本文提出了一种融合蚁群算法和随机森林的特征选择方法。该方法依据随机森林内置的变量重要性(基尼指数)评分对数据特征进行排序,将得到的特征重要性排序作为蚁群算法的启发式信息,利用蚁群算法的启发式全局优化特点对高维数据进行智能搜索,并将特征选择的结果反馈给蚁群进行信息素的实时更新。实验表明本文方法能在降低特征选择数目的同时提高数据分类的准确率。

2 研究背景

在 20 世纪 60 年代,Lewis 是最早讨论特征选择问题的,当时研究的重点是统计学和信号处理。1992 年,Kira 和 Rendell 提出 Relief 特征选择算法及其改进版本,能够选取权重比较大的特征,其余特征将被排除。90 年代后,特征选择算法中的最优特征子集被证明是 NP-hard 问题,但实际应用中可以采用启发式搜索算法,在运算效率和特征子集质量间找到一个好的平衡点[1]。

特征选择技术是数据降维的一种重要方法,它的本质就是从原始数据最初的特征集合中,选取一组符合某种评定标准的最优的特征子集,以便在进行分类或回归等任务时,可以获得更好的模型,取得更加精确的分析结果[2]。根据特征子集的选择策略,特征选择可以分为过滤式(Filter)和封装式(Wrapper)两种方法[3]。

Filter 方法先对数据集进行特征选择,然后再训练学习

本文受山东省科技发展计划项目(2014GNC110012)资助。

李光华(1996—),女,主要研究方向为人工智能、大数据,E-mail:18864802698@163.com;**李俊清**(1984—),男,硕士,副教授,主要研究方向为人工智能、大数据,E-mail:a397858801@126.com(通信作者)。

器，特征选择过程与后续学习器无关，该类方法特征选择效率较高，但对噪声数据敏感，在实际应用中一般采用此类方法进行特征的初步筛选。Wrapper 方法在特征筛选过程中直接用所选特征子集来训练分类器，之后用这个分类器在验证集上的表现评价所选的特征，该方法有很好的分类准确率，但时间复杂度高，不适用于高维数据[4]。

针对传统特征选择方法存在的问题，本文提出了一种融合蚁群算法和随机森林的特征选择方法，并通过实验验证了本文特征选择方法相比传统特征选择方法的优越性。

3　相关理论

3.1　随机森林

随机森林（Random Forest，RF）[5] 是利用多棵决策树对样本进行训练并集成预测的一种分类器，采用 Bootstrap 重抽样技术从原始样本中随机抽取数据构造多个样本，然后对每个重抽样样本采用节点的随机分裂技术构造多棵决策树，最后将多棵决策树组合，并通过投票得出最终预测结果[6]。研究证明 RF 具有很高的预测准确率，对异常值和噪声有很强的容忍度，适合处理高维数据，并且能在分析数据的同时给出变量重要性评分（Variable Importance Measures，VIM）。

3.2　特征重要度评估

本文方法采用随机森林的变量重要度评分来对特征进行评估，基本思想是：计算每个特征对随机森林里每棵树做的贡献值后取平均值，对特征之间的贡献值进行比较、排序。特征对每棵树的贡献通常可以用基尼指数（Gini index）或者袋外数据（OOB）错误率作为评价指标来衡量[7]。本文方法选择用基尼指数度量特征重要性，具体如下：

将特征重要度评分用 VIM 来表示，Gini 指数用 GI 来表示。假设有 m 个特征 $X_1, X_2, X_3, \cdots, X_m$，计算出每个特征 X_j 的 Gini 指数评分 VIM_j^{Gini}，即第 j 个特征在 RF 所有决策树中节点分裂不纯度的平均改变量，Gini 指数的计算公式为：

$$GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \tag{1}$$

其中，K 表示有 K 个类别，p_{mk} 表示节点 m 中类别 k 所占的比例，即任意从节点 m 中随机抽取两个样本，其类别标记不一致的概率。

特征 X_j 在节点 m 的重要性，即节点 m 分枝前后的 Gini 指数变化量为：

$$VIM_{jm}^{Gini} = GI_m - GI_l - GI_r \tag{2}$$

其中，GI_l 和 GI_r 分别表示分枝后两个新节点的 Gini 指数。

最后，把所有求得的重要性评分做归一化处理：

$$VIM_j^{Gini} = \frac{VIM_j^{Gini}}{\sum_{i=1}^{c} VIM_i^{Gini}} \tag{3}$$

其中，$\sum_{i=1}^{c} VIM_i^{Gini}$ 是所有特征的增益之和，VIM_j^{Gini} 是特征 x_j 的基尼指数。

3.3　蚁群算法

蚁群算法（Ant Colony Optimization，ACO）最早由意大利学者 Colorni 和 Dorigo 等于 1991 年提出[8]。它作为一种仿生学算法，是受自然界中蚂蚁觅食行为的启发而提出的。在蚂蚁觅食的过程中，每只蚂蚁在未知食物具体位置的前提下开始寻找食物，蚂蚁在前进时，会在自己走过的路径上释放出一种称为信息素的化学物质，其他蚂蚁根据不同路径上信息素的浓度来选择自己的线路，并在自己选择的路径上释放更多的信息素。某路径上走过的蚂蚁越多，信息素就越浓，该路径便会吸引更多的蚂蚁，蚂蚁之间正是通过这种信息素来合作，信息素浓度越高，对应路径被选择的概率便越高[9]。

蚁群算法能够对离散优化问题的解空间进行多点非确定性搜索，而特征选择问题本质上是一个离散组合优化问题[10]。

4　融合蚁群算法和随机森林的特征选择方法

针对传统特征选择算法对高维数据特征选择效率低、对噪声数据敏感等问题，本文提出了一种融合蚁群算法和随机森林来求解特征选择问题的方法。

4.1　问题解的表示和建立

结合蚁群算法，本文采用二进制形式表示特征选择问题。蚂蚁觅食前（见图 1），设一组特征为 $\{a_1, a_2, a_3, \cdots, a_{n-1}, a_n\}$，要从中选择一组最优特征子集。每一个特征有两种状态：一种是被选中状态，标记为 1；另一种是未被选中状态，标记为 0[12]。每只蚂蚁必须从蚁巢出发经过每一个节点并选择一条路径最终找到食物。

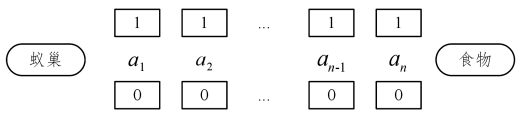

图 1　蚂蚁觅食前的特征图

蚂蚁找到食物后（见图 2），设数据集中有 5 个特征 $\{a_1, a_2, a_3, a_4, a_5\}$，蚂蚁选择的路径为 $\{1,0,0,1,1\}$，表示特征 a_1，a_4 和 a_5 被选取，特征 a_2 和 a_3 未被选取。m 只蚂蚁各自遍历一次后可得到 m 个特征子集，从中选择出最优的特征子集[9]。

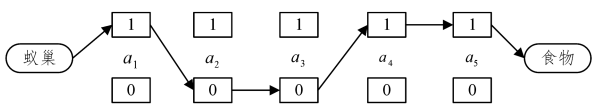

图 2　蚂蚁找到食物后的特征图

4.2　参数设置

上述模型中有几个问题需要解决：信息素初始值的设定及更新问题、路径被选择的概率设置问题以及算法的终止问题。

本文模型中，设初始时刻有 m 只蚂蚁在蚁巢中，各条路径上的信息素 τ_0 初始值相等（τ_0 可以取任意大于 0 的数值，本文方法设定 τ_0 取 0.1），每个特征共有两种信息素 τ_{ij}，i 表示特征号，j 表示特征是否被选择，被选择取值为 1，未被选择取值为 0。

其次，蚂蚁 $k(k=1,2,\cdots,m)$ 按照随机比例规则选择是否要选择下一个特征，选择概率为：

$$p_{ij}^k = \begin{cases} \dfrac{[\tau(i,j)]^{\alpha} \cdot [\eta(i,j)]^{\beta}}{\sum_{s \in J_k(i)} [\tau(i,s)]^{\alpha} \cdot [\eta(i,s)]^{\beta}}, & j \in allowed_k \\ 0, & \text{其他} \end{cases} \tag{4}$$

其中，τ_{ij} 是特征 i 取 j 值的信息素；α 是信息素重要程度的参数；$\eta = \dfrac{1}{d_{ij}}$ 是特征 i 取 j 的启发式信息，其中 $d_{ij} = \dfrac{1}{VIM_{ij}}$，$VIM_{ij}$ 求解见式(2)，即随机森林特征重要度评分；β 是启发式信息重要程度的参数；$allowed_k$ 是蚂蚁 k 下一步要访问的特征集合。

经过 t 时刻，所有蚂蚁完成一次遍历后要更新各边上的

信息素，首先是信息素的挥发，其次是蚂蚁在它们所经过边上释放信息素，更新公式如下：

$$\tau_{ij}(t)=(1-\rho)\tau_{ij}(t-1) \tag{5}$$

其中，ρ 为信息素挥发系数，通常设置 $0<\rho\leqslant 1$。

$$\tau_{ij}(t)=\tau_{ij}(t-1)+\sum_{k=1}^{m}\Delta\tau_{ij}^{k} \tag{6}$$

其中，$\Delta\tau_{ij}^{k}$ 是第 k 只蚂蚁向它经过的边释放的信息素，定义为：

$$\Delta\tau_{ij}^{k}=\begin{cases}\dfrac{1}{d_{ij}}, & \text{如果边 } (i,j) \text{在路径 } T^{k} \text{ 上}\\ 0, & \text{否则}\end{cases} \tag{7}$$

根据式(7)可知，蚂蚁构建的 d_{ij} 越小，即特征重要性评分越大，路径上各条边就会留下更多的信息素，该路径在之后的迭代中被其他蚂蚁选择的概率就越高。

算法终止条件可以是得到最优解即经过连续迭代没有进一步提高或者达到设定的最大运行次数，本文方法选择达到最大运行次数为终止条件。

4.3 算法描述

本文提出蚁群算法(AOC)和随机森林(RF)融合的特征选择方法——融合蚁群算法和随机森林的特征选择方法(简记为 AOC_RF)。图 3 为 AOC_RF 算法的流程图。

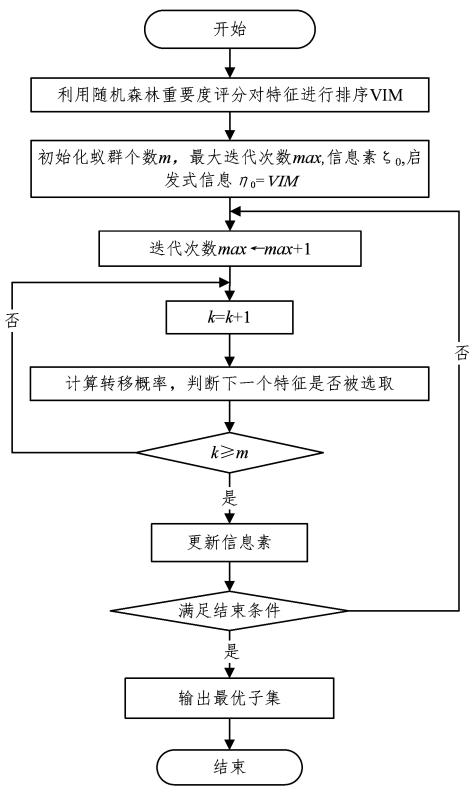

图 3 AOC_RF 流程图

Step1 按式(2)计算随机森林特征重要度评分 VIM，将高维数据的特征进行排序。

Step2 初始化蚁群算法的参数，包括蚂蚁个数 m、迭代次数 max、初始信息素水平 τ_0、启发式信息 $\eta_0=$VIM。

Step3 初始化问题空间，让蚁群算法对高维数据进行智能搜索。

Step4 按式(4)计算蚂蚁在节点之间转移的概率，构造特征子集的解空间。

Step5 将特征选择的结果作为评价器，根据式(6)不断更新信息素。

Step6 若迭代次数小于最大迭代次数，返回 Step4；否则，终止计算，输出最优的特征子集。

5 实验结果及分析

实验数据均来自 UCI[13] 机器学习库中特征数目较大(多于 30)的数据，分别为 Dermatology，Handwrite，Ionosphere，WDBC，WPBC。实验数据集描述如表 1 所列。

表 1 选择测试的 UCI 数据集描述

序号	数据集	样本个数	特征数目	分类数目
1	Dermatology	358	34	6
2	Handwrite	323	255	2
3	Ionosphere	351	34	2
4	WDBC	569	30	2
5	WPBC	184	33	2

本文实验环境为 DELL Inspiron 15 5000 Series 笔记本，Intel(R) Core(TM) i5-5200 2.20 GHz CPU，4.0 GB 内存，Windows10 64 位操作系统，软件环境为 PyCharm 2018.3.2、Python 3.6.3。

在分类准确率方面，将随机森林作为分类器，对本文方法筛选的特征、ReliefF[14] 算法筛选的特征和未筛选的特征进行了比较；在选择的特征数目方面，将本文方法筛选的特征数目和 ReliefF 算法筛选的特征数目进行了比较。

表 2 是本文方法与原始数据在随机森林上分类准确率的比较结果。实验进行 10 次，表中结果用计算得到的 10 次分类准确率的平均值百分比±标准差来表示。

表 2 本文方法和原始数据的分类准确率比较

(单位：%)

数据集	本文方法	原始数据
Dermatology	95.67±2.33	93.79±1.31
Handwrite	97.28±0.92	97.07±0.36
Ionosphere	97.65±0.66	93.96±0.91
WDBC	96.00±0.23	96.54±0.93
WPBC	90.48±3.36	76.78±1.39

如表 2 所列，在 WDBC 数据集上，本文方法分类准确率(96.00±0.23)%总体与原始数据的分类准确率(96.54±0.93)%持平，但本文方法的标准差小于原始数据，即本文方法的分类稳定性更高；在其余 4 个数据集上，用本文方法筛选过的特征进行分类的性能都优于原始数据，分类准确率都有不同程度的提升。

表 3 本文方法和 ReliefF 算法的实验结果

数据集	特征数	本文方法		ReliefF 算法	
		分类准确率/%	选择的特征数	分类准确率/%	选择的特征数
Dermatology	34	96.11±1.82	12	87.56±0.60	15
Handwrite	255	97.11±1.49	10	89.77±2.81	12
Ionosphere	34	97.46±0.59	8	90.14±0.93	10
WDBC	30	96.00±0.23	5	94.10±0.61	7
WPBC	33	92.14±2.25	3	88.57±3.68	4

表 3 为本文的特征选择方法与 ReliefF 算法的比较结果。与 ReliefF 算法相比，本文方法在分类准确率和选择特征数目上的性能都有明显的改善。在 Dermatology 数据集上，本文方法分类准确率的标准差略高于 ReliefF 算法；但本文方法的特征分类准确率为(96.11±1.82)%，总体高于传统的 ReliefF 算法的分类准确率(87.56±0.60)%；在 Handwrite，Ionosphere，WDBC 和 WPBC 数据集上，本文方法分类准确率

均明显高于 ReliefF 算法的分类准确率且本文选取的特征数明显少于 ReliefF 算法选取的特征数。

综上所述，本文方法在分类准确率、分类稳定性上均优于传统的特征选择算法，且特征选择数目较少，达到了高维空间维度约简的目标。

结束语　大数据时代，如何降低高维空间中特征选择的计算成本并提高分类准确率是数据挖掘领域中的重要研究问题之一。本文在分析传统特征选择算法的基础上，提出了一种融合蚁群算法和随机森林的特征选择方法，并在 UCI 数据集上将两种方法进行了对比，实验表明本文的特征选择方法能够有效地减少数据集中的特征数量，同时提高了数据分类的准确率。

参 考 文 献

[1] 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(工学版)，2014，44(1)：137-141.

[2] 刘飞飞. 特征选择算法及应用综述[J]. 办公自动化，2018，23(21)：47-49.

[3] 张翠军，陈贝贝，周冲，等. 基于多目标骨架粒子群优化的特征选择算法[J]. 计算机应用，2018，38(11)：3156-3160，3166.

[4] 刘依恋. 模式分类中特征选择算法研究[D]. 哈尔滨：哈尔滨理工大学，2014.

[5] BREIMEN L. Random Forests [J]. Machine Learning，2001，45(1)：5-32.

[6] 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.

[7] 杨凯，侯艳，李康. 随机森林变量重要性评分及其研究进展[OL]. http://www.paper.edu.cn/releasepaper/content/201507-212.

[8] ALBERTO C，MANIEZZO D. Distributed optimization by ant colonies[C]// Proc of the First European Conf on Artificial Life. Paris：Elsevier Publishing. 1991：134-142.

[9] 黄丹凤，祁云嵩，许姗娜. 基于粗糙集和蚁群算法的特征基因选择方法[J]. 计算机技术与发展，2012，22(6)：68-70，74.

[10] 马军建，董增川，王春霞，等. 蚁群算法研究进展[J]. 河海大学学报(自然科学版)，2005(2)：139-143.

[11] 杨丽. 基于 ReliefF 和蚁群算法的特征基因选择方法分析[J]. 电脑知识与技术，2017，13(32)：199-200.

[12] MURPHY P M，AHA D W. UCI repository of machine learning database [DB/OL]. (2006-05-12). http://www.ics.uci.edu/mlearn/MLRepository.html.

[13] KIRA K，RENDELL L A. The feature selection problem：Traditional methods and a new algorithm[C]// AAAI. 1992：129-134.

[14] 卜华龙，夏静，韩俊波. 特征选择算法综述及进展研究[J]. 巢湖学院学报，2008(6)：41-44.

[15] 许行，张凯，王文剑. 一种小样本数据的特征选择方法[J]. 计算机研究与发展，2018，55(10)：2321-2330.

[16] 朱振国，赵凯旋，刘民康. 基于强化学习的特征选择算法[J]. 计算机系统应用，2018，27(10)：214-218.

[17] 闫春，李亚琪，孙海棠. 基于蚁群算法优化随机森林模型的汽车保险欺诈识别研究[J]. 保险研究，2017(6)：114-127.

[19] 雷海锐，高秀峰，刘辉. 基于机器学习的混合式特征选择算法[J]. 电子测量技术，2018，41(16)：42-46.

[20] 邱宁佳，周稳，王鹏，等. 一种结合改进 CHI 和 RFFS 的特征选择算法研究[J]. 计算机工程与应用，2018，54(21)：133-140.

[21] 叶志伟，郑肇葆，万幼川，等. 基于蚁群优化的特征选择新方法[J]. 武汉大学学报(信息科学版)，2007(12)：1127-1130.

[22] 张文杰，蒋烈辉. 一种基于遗传算法优化的大数据特征选择方法[J]. 计算机应用研究，2019：1-5.

[23] 李晓岚. 基于 Relief 特征选择算法的研究与应用[D]. 大连：大连理工大学，2013.

[24] 蔡萌萌，张巍巍，王泓霖. 大数据时代的数据挖掘综述[J]. 价值工程，2019，38(5)：155-157.

[25] 魏茂胜. 数据挖掘中的分类算法综述[J]. 网络安全技术与应用，2017(6)：65-66.

(上接第 207 页)

[10] ZHAO Q，SHI Y，LIU Q，et al. A grid-growing clustering algorithm for geospatial Data[J]. Pattern Recognition Letters，2014，53(53)：77-84.

[11] KUMAR K M，REDDY A R M. A fast DBSC-AN clustering algorithm by accelerating neighbor searching using Groups method[J]. Pattern Recognition，2016，58：39-48.

[12] HE Y，TAN H，LUO W，et al. MR-DBSCAN：An efficient parallel density-based clustering algorithm using MapReduce[C]// 2011 IEEE 17th International Conference on Parallel and Distributed Systems. IEEE Computer Society，2011：473-480.

[13] KIM Y，SHIM K，KIM M S，et al. DBCURE-MR：An efficient density-based clustering algorithm for large data using MapReduce[J]. Information Systems，2014，42(2)：15-35.

[14] 于彦伟，贾召飞，曹磊，等. 面向位置大数据的快速密度聚类算法[J]. 软件学报，2018，29(8)：2470-2484.

[15] ESTER M，KRIEGEL H P，SANDER J，et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery & Data Mining. Portland：AAAI Press，1996：226-231.

[16] 黄德才. 数据仓库与数据挖掘教程[M]. 北京：清华大学出版社，2016.

[17] 钱潮恺，黄德才. 基于维度频率相异度和强连通融合的混合数据聚类算法[J]. 模式识别与人工智能，2016，29(1)：82-89.

[18] 余长俊，张燃. 云环境下基于 Canopy 聚类的 FCM 算法研究[J]. 计算机科学，2014，41(S1)：316-319.

[19] GIONIS A，MANNILA H，TSAPARAS P. Clustering aggregation[J]. ACM Transactions on Knowledge Discovery from Data (TKDD)，2007，1(1)：1-30.

[20] ZAHN C T. Graph-theoretical methods for detecting and describing gestalt clusters[J]. IEEE Transactions on Computers，1971，100(1)：68-86.

[21] YUAN J，ZHENG Y，XIE X，et al. T-Drive：Enhancing driving directions with taxi drivers' intelligence[J]. IEEE Transactions on Knowledge & Data Engineering，2013，25(1)：220-232.