Computer Science and Application 计算机科学与应用, 2023, 13(4), 883-889
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134086

Hans汉斯

基于XGBoost和蚁群算法的特征选择方法

张凌翱

青岛科技大学信息科学技术学院，山东 青岛

收稿日期：2023年3月22日；录用日期：2023年4月21日；发布日期：2023年4月28日

摘　要

在机器学习领域，处理高维特征数据时通常会面临冗余和不相关的特征问题，因此特征选择成为一个重要的挑战。对于多维度数据，Relief算法作为一种传统的特征选择算法，具有较高的计算效率和较好的稳定性，被大量应用于实际场景，但Relief算法的特征选择结果具有随机性，不同的初始采样会有不同的结果，且对于特征之间存在较强依赖关系的数据集，如共线性等，可能会导致结果不准确。本文提出了一种特征选择方法，称为X-ACO方法，它结合了XGBoost和蚁群算法。本文算法蚁群路径搜索过程的启发式信息使用XGBoost算法的特征重要性来表示。同时，使用特征之间的皮尔森相关系数来调整信息素浓度，以便更好地控制特征的相关性。实验证明，X-ACO方法可以在保证分类准确率的前提下，减少特征数量，降低特征冗余，并提高算法性能。

关键词

特征选择，XGBoost，蚁群算法，皮尔森系数

Feature Selection Method Based on XGBoost and Ant Colony Optimization

Ling'ao Zhang

School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao Shandong

Received: Mar. 22nd, 2023; accepted: Apr. 21st, 2023; published: Apr. 28th, 2023

Abstract

In the field of machine learning, the problem of redundant and irrelevant features is usually faced when dealing with high-dimensional feature data, so feature selection becomes an important challenge. For high-dimensional data, Relief algorithm, as a commonly used feature selection algorithm,

文章引用: 张凌翱. 基于 XGBoost 和蚁群算法的特征选择方法[J]. 计算机科学与应用, 2023, 13(4): 883-889.
DOI: 10.12677/csa.2023.134086

has high computational efficiency and good stability, and is heavily used in practical scenarios, but the feature selection results of Relief algorithm have randomness, different initial sampling will have different results, and it may lead to inaccurate results for data sets with strong dependencies between features, such as covariance. In this paper, we propose a feature selection method, called X-ACO method, which combines XGBoost and ant colony optimization. The method uses the feature importance of the XGBoost algorithm as heuristic information for the ant colony path search process of the algorithm in this paper. Meanwhile, the Pearson correlation coefficient between features is used to adjust the pheromone concentration in order to better control the relevance of features. Experiments demonstrate that the X-ACO method can reduce the number of features, reduce feature redundancy, and improve the algorithm performance while ensuring classification accuracy.

Keywords

Feature Selection, XGBoost, Ant Colony Optimization (ACO), Pearson Coefficient

1. 引言

在机器学习处理实际问题中，往往存在着大量冗余、无关和不必要的特征[1]，这些特征会占用大量的存储空间、计算资源和时间成本，同时还会导致模型过拟合和泛化能力下降。因此，特征筛选[2] [3]已经成为了机器学习领域的重点研究问题之一。通过合理的特征选择，可以减少特征维度，提高模型的可解释性和泛化能力，从而更好地解决实际问题。本文提出了一种特征选择方法(X-ACO 方法)，将XGBoost 和蚁群算法结合在一起，根据 XGBoost 的内置特征重要性评分进行特征排序[4]，同时利用特征间的 Pearson 相关系数表示特征间的距离，根据特征相关性系数调节信息素浓度，使用各特征重要性作为本文蚁群算法部分的启发式函数。实验表明，这种方法可以确保选取的特征之间不会产生强烈的线性关联，减少特征冗余，提高数据分类的精确性，同时还可以减少选择的特征数量。

2. 研究背景

在实际应用中，特征选择在数据预处理过程中非常重要，可以帮助去除冗余和无关特征，提高模型的解释性和泛化性能。特征选择的研究已经有很长的历史，并且一直受到学术界和工业界的广泛关注。近年来，随着机器学习算法和数据科学应用的广泛应用，特征选择的重要性也越来越突出。然而，由于现实数据集通常包含大量特征，因此进行特征选择时往往会面临一些挑战。例如，可能存在高度相关的特征，这会导致模型过度拟合和性能下降。另外，某些特征可能会包含噪声或无关信息，这也会影响模型的精度和可解释性。

为了应对这些挑战，学术界和工业界已经提出了许多特征选择方法和算法[5] [6] [7] [8] [9]。这些方法可以根据不同的目标和需求，选择最佳的特征子集，以提高模型的性能和效率。一些常用的特征选择方法包括基于过滤器的方法、基于包装器的方法和基于嵌入式的方法等。Relief 算法就是一种常用的特征选择算法，具有较高的计算效率和较好的稳定性，但 Relief 算法的特征选择结果具有随机性，不同的初始采样会有不同的结果，且对于特征之间存在较强依赖关系的数据集，如共线性等，可能会导致结果不

准确。

为了解决上述问题，本文提出了一种将 XGBoost 与蚁群算法相结合的特征选择方法，并引入了皮尔森(Pearson)系数来表示特征间的相关性系数，经实验验证，相对于传统特征选择方法，本文特征选择方法可以更好地改善模型的各种性能与泛化能力。

3. 相关理论

3.1. XGBoost 算法

梯度提升决策树(Extreme Gradient Boosting, XGBoost)即一种高效的梯度提升决策树算法。它在原有的 GBDT 基础上进行了改进，使得模型效果得到大大提升。作为一种前向加法模型，他的核心是采用集成思想——Boosting 思想，将多个弱学习器通过整合为一个强学习器。即用多棵树共同决策，并且用每棵树的结果都是目标值与之前所有树的预测结果之差，并累加所有的结果，得到最终的结果，以此达到整个模型效果的提升[10] [11] [12] [13]。

XGBoost 是由多棵 CART (Classification And Regression Tree)，即分类回归树组成，因此它可以处理分类回归等问题。研究证明，在高维数据处理方面，XGBoost 算法可以根据特征重要性自动选择特征，提高模型的泛化能力，降低过拟合。

3.2. XGBoost 特征重要度

本文方法选择 XGBoost 算法的基尼系数来度量特征重要性，具体原理公式如下：

假设有 m 个特征 $X_1, X_2, X_3, \cdots, X_m$，用 V 表示特征重要性，计算出每个特征 X_j 的 Gini 指数评分 V_j^{Gini}。Gini 指数的计算公式为：

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

V_{jm}^{Gini} 表示特征 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：

$$V_{jm}^{\text{Gini}} = G_m - G_l - G_r \tag{2}$$

对求得的各特征的重要性进行归一化：

$$V_j^{\text{Gini}} = \frac{V_j^{\text{Gini}}}{\sum_{i=1}^{c} V_i^{\text{Gini}}} \tag{3}$$

其中，V_j^{Gini} 是特征 X_j 的基尼指数，$\sum_{i=1}^{c} V_i^{\text{Gini}}$ 是特征的增益之和。

3.3. 皮尔森(Pearson)系数

皮尔森相关系数即皮尔森积矩相关系数(Pearson Product-moment Correlation Coefficient)，是最常用的一种线性相关系数。用来反映两个属性 X 和 Y 的线性相关程度，绝对值越大表明属性间相关性越强。

两个属性间的皮尔森系数越高，属性间的相关性越强，在特征选择时，尽量避免同时选取相关性较强的特征。皮尔森系数的计算公式如下：

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{4}$$

其中，$\operatorname{cov}(X,Y)$ 表示属性 X，Y 的协方差，$\sigma_X \sigma_Y$ 表示属性 X，Y 的标准差的乘积。

3.4. 蚁群算法

蚁群算法(Ant Colony Optimization, ACO)由意大利学者 Colorni 等人于 1991 年提出[14]，是一种基于蚂蚁觅食行为的启发式优化算法。该算法模拟了蚂蚁觅食的过程：每只蚂蚁在寻找食物的过程中，会释放一种化学物质——信息素，且能够感知到其它蚂蚁释放的信息素。蚂蚁之间正是通过这种信息素来选择觅食路线，信息素浓度越高，对应路线被选择的概率便越高[15]。

蚁群算法通过信息素的挥发与更新来搜索最优解，常用于解决优化问题[16]，如最小化函数值，寻找最优路径等。

4. 基于 XGBoost 和蚁群算法的特征选择方法

针对传统 Relief 选择算法对多维度数据特征选择有效性差、对存在较强依赖关系的特征选择结果不精确等问题，本文提出了一种基于 XGBoost 和蚁群算法的特征选择方法，简称 X-ACO。

4.1. 路径转移概率计算

对于每一只蚂蚁来说，它经过的每一个特征，都有两种状态：0 或 1，0 表示该节点未被选中状态，1 表示被选中状态。对于每只蚂蚁从初始位置出发到终点的路径上，对于特征 i 和 j，则存在 4 条可供选择的路径，分别是 0→0、0→1、1→0、1→1。

对于每只蚂蚁所经过路径上的每一个节点，其被选择的概率如下：

$$p_{i,j}^{k}(t)=\begin{cases}\dfrac{\tau_{i,j}^{\alpha}\eta_{i,j}^{\beta}}{\sum_{k\in C}\tau_{i,k}^{\alpha}\eta_{i,k}^{\beta}}, & j\in C\\ 0, \text{ 否则}\end{cases} \tag{5}$$

其中，i 表示当前节点；j 表示下一个待选节点；C 表示蚂蚁能够访问且尚未被访问到的节点集合；$\tau_{i,j}$ 表示节点间的信息素；α 是信息素因子；$\eta_{i,j}$ 表示节点的启发性信息；β 是启发式信息因子。

4.2. 参数设置与信息素的更新

设初始时刻有 n 只蚂蚁，并随机初始化 n 只蚂蚁的初始位置，各条路径上的信息素 τ_0 初始值相等。$\tau_{i,j}=V$，XGBoost 特征重要性 V 的计算公式见式(2)，经过 t 时刻，当蚁群的所有蚂蚁全部完成一次遍历后，及时更新各条路径上的信息素[13]，首先是各路径上信息素的挥发，更新公式如下：

$$\tau_{i,j}(t)=(1-\rho)*\tau_{i,j}(t-1) \tag{6}$$

其中，ρ 为信息素挥发系数，通常设置 $0<\rho\le 1$。其次是蚂蚁在它们所经过边上释放信息素。

$$\tau_{i,j}(t)=\tau_{i,j}(t-1)+\sum_{k=1}^{n}\Delta\tau_{i,j}^{k} \tag{7}$$

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

$$\Delta\tau_{i,j}^{k}=\begin{cases}\dfrac{Q}{\rho_{i,j}}, & \text{边}(i,j)\text{在路径上}\\ 0, \text{ 否则}\end{cases} \tag{8}$$

其中，Q 为信息素常数，表示每只蚂蚁循环一次释放的信息素总量，$\rho_{i,j}$ 为路径长度，即皮尔森相关系数，见公式(4)。结合公式(5)和公式(8)可知，皮尔森线性相关系数越小，该路径上的信息素浓度越高，且特征重要度越高，该路径被其他的蚂蚁选择的概率就越高。

4.3. 算法描述

本文结合 XGBoost 算法和蚁群算法，并引入皮尔森系数作为节点间的路径长度，提出一种特征选择方法，X-ACO 算法。算法描述如下：首先运用 XGBoost 算法来计算每个特征的特征重要度 V；然后利用皮尔森系数计算函数，见式(4)，算出特征间的相关性系数 ρ；接下来构建 X-ACO 算法，计算各特征的选择概率 P，并进行参数的初始化和信息素 τ 的更新；最后运用 X-ACO 方法输出选择的最优特征子集 S_x。

为了对本文算法进行更为准确直观的描述，现给出 X-ACO 算法伪代码见表 1。

Table 1. X-ACO algorithm pseudo-code
表 1. X-ACO 算法伪代码

输入：数据集 S

输出：最优特征子集 S_x

1. 计算数据集各特征重要性 V。

2. 构建 X-ACO 算法：

a. 初始化信息素 τ_0，蚂蚁个数 n，启发性函数 V，蚂蚁随机初始位置，最大迭代次数 max。

b. 运用式(4)计算特征间的相关性系数。

c. for i =1 to max do

for i =1 to m do

运用式(5)计算每个蚂蚁在每条路径上的选择概率 P。

End for

运用式(6)、式(7)、式(8)更新信息素 τ 。

End for

d. 输出最优特征子集 S_x。

5. 实验结果与分析

本文从 UCI 数据库中选择了一些经典的数据集对算法进行验证。经典的数据集见表 2。

Table 2. UCI dataset description
表 2. UCI 数据集描述

数据集	样本数	特征数	分类数
wpbc	198	33	2
wdbc	569	30	2
Sonar	208	60	2
Spambase	4601	57	2
Wine	178	13	3

5.1. 实验环境与方法

本文实验环境为 ThinkPad E550c 笔记本，Intel(R) Core(TM) i3-4005U CPU @ 1.70 GHz， 4 .0 GB 内存，Windows 10 64 位操作系统，软件环境为 Jupyter 6.3.0、Python 3.8.3。

在分类器选择上，选取逻辑回归作为分类器，对本文方法选取的特征、与原始数据特征进行了分类预测准确率比较；首先对实验数据进行预处理，先将各数据进行归一化操作，将每个特征属性列的数据都归一到[0, 1]之间，然后对预处理后的数据进行特征选择实验。本文方法与数据集原始特征数据在逻辑回归分类器上的分类准确率和特征个数的对比结果见表 3。

5.2. 实验分析

对比分类精度，从表 3 不难看出，在 UCI 数据集上，本文方法分类准确率相较于原始数据的分类准确率均取得较优提升。本文方法选取特征更少，拟合效果较优。通过对比表 4 中的精确率、F1 分数以及 AUC 三项评价指标发现 X-ACO 算法大多数情况下三项评价指标都略高于 Relief 算法。

Table 3. Comparison of classification accuracy and number of features
表 3. 分类准确率和特征个数对比

数据集	本文方法	原始数据	本文选取特征个数	原始特征数
wpbc	0.929	0.901	3	33
wdbc	0.952	0.936	5	30
Sonar	0.807	0.692	6	60
Spambase	0.925	0.894	5	57
Wine	0.981	0.944	3	13

Table 4. Experimental performance comparison between X-ACO and Relief algorithms
表 4. X-ACO 算法与传统 Relief 算法实验性能比较

算法	指标	wpbc	wdbc	Sonar	Spambase	Wine
Relief	ACC	0.918	0.941	0.814	0.880	0.962
	F1	0.892	0.912	0.826	0.856	0.956
	AUC	0.902	0.933	0.765	0.877	0.952
X-ACO	ACC	0.929	0.952	0.807	0.933	0.981
	F1	0.905	0.930	0.852	0.918	0.981
	AUC	0.924	0.945	0.781	0.929	1

表 4 为本文方法 X-ACO 与传统 Relief 特征选择算法在 UCI 的数据集上各项评价指标对比结果。与 Relief 算法相比，本文方法在分类准确率、F1 分数、AUC 等指标性能上都有明显的优势。在 Sonar 数据集上，本文方法分类准确率略低于 Relief 算法；但本文方法的 F1 分数为 0.852，高于 Relief 算法的 F1 分数 0.826；在 wpbc、wdbc 和 Spambase 数据集上，本文方法分类准确率、F1 分数、AUC 均优于传统 Relief；在多分类数据集 Wine 上，本文方法在分类准确率、加权 F1 分数、AUC 三项评价指标上均优于 Relief 算法。

综上所述，本文 X-ACO 特征选择方法在分类准确率、F1 分数、AUC 上均优于传统的特征选择算法，不仅达到了数据降维的目标，同时提高了算法的泛化能力。

6. 结论

本文提出的基于 XGBoost 算法和蚁群算法的特征选择方法(X-ACO 方法)，运用各个特征的 XGBoost

特征重要度作为蚁群算法部分的启发式函数，同时利用特征间的皮尔森(Pearson)相关系数表示各特征之间的距离长短，进而选择最优特征子集，从而达到数据降维的效果。实验表明，本文方法能在降低特征维度的同时，确保数据集分类的准确率，减少特征冗余，提高算法的各项性能，为实际生产中的特征选择问题提供了重要的方法依据。

参考文献

[1] 孙洁丽, 刘沛, 翟浩文. 基于高维数据的聚类研究综述[J]. 河北省科学院报, 2022, 39(5): 1-6.

[2] 邹丽英, 刘祎. 超高维缺失响应数据的特征筛选[J]. 中国海洋大学学报(自然科学版), 2023, 53(1): 147-156.

[3] 钟彩, 杨亚鑫, 王璟德, 孙巍. 特征筛选对抗肿瘤药物识别的影响研究[J]. 化学研究与应用, 2022, 34(10): 2350-2356.

[4] 罗妍, 王枞, 叶文玲. 基于 XGBoost 和 SHAP 的急性肾损伤可解释预测模型[J]. 电子与信息学报, 2022, 44(1): 27-38.

[5] 熊玲珠, 邱伟涵, 罗计根, 李科定. 基于最大信息系数和迭代式XGBoost的混合特征选择方法[J]. 计算机应用与软件, 2023, 40(1): 280-286+305.

[6] 徐久成, 孟祥茹, 瞿康林, 孙元豪, 杨杰. 基于模糊邻域相对依赖互信息的特征选择方法[J]. 模糊系统与数学, 2023, 37(1): 121-135.

[7] 何鹏, 龙文. 一种改进鲸鱼优化算法的特征选择方法[J]. 绿色科技, 2022, 24(18): 246-248+271.

[8] 孙林, 施恩惠, 司珊珊, 徐久成. 基于 AP 聚类和互信息的弱标记特征选择方法[J]. 南京师大学报(自然科学版), 2022, 45(3): 108-115.

[9] Abdulhussien, A.A., Nasrudin, M.F., Darwish, S.M. and Abdi, A.A.Z. (2023) Feature Selection Method Based on Quantum Inspired Genetic Algorithm for Arabic Signature Verification. *Journal of King Saud University—Computer and Information Sciences*, **35**, 141-156. https://doi.org/10.1016/j.jksuci.2023.02.005

[10] 刘江, 许康智, 蔡伯根, 郭忠斌, 王剑. 基于 XGBoost 的列控车载设备故障预测方法[J]. 北京交通大学学报, 2021, 45(4): 95-106.

[11] Suenaga, D., Takase, Y., Abe, T., Orita, G. and Ando, S. (2023) Prediction Accuracy of Random Forest, XGBoost, LightGBM, and Artificial Neural Network for Shear Resistance of Post-Installed Anchors. *Structures*, **50**, 1252-1263. https://doi.org/10.1016/j.istruc.2023.02.066

[12] Ali, S., Khorrami, B., Jehanzaib, M., *et al*. (2023) Spatial Downscaling of GRACE Data Based on XGBoost Model for Improved Understanding of Hydrological Droughts in the Indus Basin Irrigation System (IBIS). *Remote Sensing*, **15**, Article 873. https://doi.org/10.3390/rs15040873

[13] Ren, Q.X. and Wang, J.G. (2023) Research on Enterprise Digital-Level Classification Based on XGBoost Model. *Sustainability*, **15**, Article 2699. https://doi.org/10.3390/su15032699

[14] Colorni, A., Dorigo, M. and Maniezzo, V. (1991) Distributed Optimization by Ant Colonies. *Proceedings of the First European Conference on Artificial Life*, Vol. 142, 134-142.

[15] 郭城成, 田立勤, 武文星. 蚁群算法在求解旅行商问题中的应用综述[J]. 计算机系统应用, 2023, 32(3): 1-14.

[16] 郭琴, 郑巧仙. 基于优化蚁群算法的机器人路径规划[J]. 湖北大学学报(自然科学版), 2023, 45(2): 157-163.