

融合人类知识的随机森林特征 选择方法研究

戴贵洋, 綦秀丽*, 余晓晗

(陆军工程大学 指挥控制工程学院, 江苏 南京 210007)

摘要:特征选择可以从原始特征空间中选择出一些最有效的特征以降低数据特征维度,提高学习算法性能。在数据降维问题中,常见的特征选择方法主要依靠数据本身的统计特性,通过数据本身信息选择更有效的特征,然而一些实际问题中往往积累了大量人类经验,这些人类知识可能对特征选择有重要影响,但很少有特征选择方法考虑使用这些人类知识。针对此类包含人类知识问题,并兼顾人类知识和采集数据的特征选择方法,提出了基于随机森林和模糊系统的二次筛选的特征选择模型。该模型通过随机森林算法剔除原始数据集中的冗余特征,实现初步筛选,利用初选特征中包含的人类知识搭建模糊系统,对初选特征计算评估得分,筛选出最终的关键特征。在汽油提纯真实数据集上进行了实验,相较于常规特征选择方法,该模型有显著提升,验证了结合人类知识随机森林特征选择方法的有效性。

关键词:特征选择;随机森林;人类知识;模糊系统;数据降维

中图分类号:TP182;TP391

文献标识码:A

文章编号:1673-629X(2022)07-0155-06

doi:10.3969/j.issn.1673-629X.2022.07.027

Research on Random Forest Feature Selection Method by Human Knowledge

DAI Gui-yang, QI Xiu-li*, YU Xiao-han

(School Command & Control Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: Feature selection methods can select more efficient features from the original feature space to reduce data characteristic dimensions and improve learning algorithm performance. For the problem of data dimensionality reduction, common feature selection methods mainly rely on the statistical characteristics of the data itself, and select more effective features through the data itself. However, a lot of human experience is often accumulated in some practical problems. Human knowledge may have an important influence on feature selection, but few feature selection methods take the use of such human knowledge into account. In response to this kind of feature selection method that contains human knowledge and takes into account both human knowledge and collected data, a feature selection model based on secondary screening of random forest and fuzzy system is proposed. The model uses the random forest algorithm to eliminate redundant features in the original data set to achieve preliminary screening, build a fuzzy system using human knowledge contained in primary elections, calculate evaluation scores for the primary selected features, and screen out the final key features. Experiments were carried out on the real data set of gasoline purification. Compared with the conventional feature selection method, the model has a significant improvement, which verifies the effectiveness of the random forest feature selection method combined with human knowledge.

Key words: feature selection; random forest; human knowledge; fuzzy system; data dimensionality reduction

0 引言

特征选择算法通过剔除冗余和不重要的特征,从原始特征空间中选择出最具分辨力的特征子集,以帮助分类器提高分类精度。主流的特征选择方法大部分

是基于机器学习模型的方法,有些机器学习方法本身就具有对特征进行打分的机制,因而很容易被运用到特征选择任务中,如回归模型、SVM、决策树和随机森林等。现实生活中,特征选择算法已经广泛地应用于

收稿日期:2021-07-16

修回日期:2021-11-17

基金项目:国家自然科学基金项目(61806221)

作者简介:戴贵洋(1995-),男(满),硕士研究生,研究方向为人工智能、数据工程;通讯作者:綦秀丽(1972-),男,硕士,教授,研究方向为计算机仿真。

各个领域,如数据挖掘^[1-3]、信息融合^[4-5]、模式识别^[6-8]等。

在大多数已有研究中,特征选择算法是基于原始数据集进行特征筛选,算法的本质更关注数据中隐含的特征之间关系。熊熙等人^[9]提出一种称为 FOAD (fuzzy-option based attribute discriminant method) 的基于模糊选项关系的关键特征选择方法,通过数据获取、模糊选项的选择与约简以及关键特征的排序与提取对每个参与者样本包含的若干特征进行筛选,为每个特征都选择一个程度选项,从而提取出关联度更高的特征。孙广路等人^[10]发现最大信息系数 (maximum information coefficient, MIC) 可以对特征变量间的线性和非线性关系,以及非函数依赖关系进行有效度量,并提出了一种评价各维特征间相关性的度量标准,基于新度量标准又提出近似马尔可夫毯特征选择方法,删除冗余特征。这些算法对数据集中包含的信息更为关注,然而在一些应用中,人类专家积累了有关特征关联关系的经验,这些经验并不一定会体现到数据集中,因此常常与数据集中的信息有较高的独立性,对特征选择帮助较大。例如,在汽油提纯过程中,通过多年积累的专家经验可知氢油比越高,辛烷值损失越少等。这些信息无法直接由一般的特征选择算法学习得到,因此应用到特征选择上时会起到意想不到的效果。由此,该文尝试搭建一个框架将人类知识引入特征选择任务中,从而筛选出更具代表性的关键特征。

在随机森林特征选择模型中,设计了引入人类知识提升特征选择效果的方法,通过基于模糊系统对人类知识建模,建立了人类主观知识和客观数据集的联系,完成了对数据集特征的二次筛选,从大量特征中更好地筛选出关键特征。通过实验验证,较单一的随机森林特征选择算法,该方法在关键特征选择问题上具有更高的可靠性。

1 汽油提纯特征选择问题描述

现实生活中,汽油燃烧产生的尾气排放是污染大气环境的重要因素,汽油清洁化工作长期受到各个国家的高度重视。为了有效利用重油资源,必须对催化裂化汽油进行精制处理,以满足对汽油质量的要求。辛烷值 (RON) 是反映汽油燃烧性能的最重要指标,某石化企业积累了大量历史数据,其中对辛烷值损失和产品硫含量有影响的特征变量包括 7 个原料性质、2 个待生吸附剂性质、2 个再生吸附剂性质、2 个产品性质等变量以及另外 354 个操作变量 (共计 367 个特征变量)。这些特征变量中,大部分特征变量对辛烷值损失及产品硫含量影响较小,少量特征变量是针对催化裂化汽油进行脱硫和降烯烃过程中的关键特征变

量,然而,各个特征变量之间又可能相互作用,相互影响,这使得特征变量之间耦合性较高。为了工程方便,提高汽油提纯工作效率,需要对 367 个特征变量进行关键特征选择。一般来说,常见的特征选择算法在处理大量耦合的非线性数据中,很难选择出少部分关键变量。

在汽油提纯过程中,一般的特征选择算法,如随机森林算法通常根据决策树的投票来选择最优的特征变量,每个特征都试图代表局部最优解,这些算法通常只关注数据集的本身,容易受到数据采样方式的影响,从而疏漏关键变量。研究发现,影响辛烷值变化除了反映在客观数据中,还会积累到专家经验中,例如,氢油比含量越高,辛烷值损失越少,产品中硫含量越低;反应器温度控制在 421℃ 左右时,辛烷值损失少,产品中硫含量低。这些专家经验往往不容易在数据采样中被捕捉,可以作为基于数据集特征选择算法的补充,提升获取关键特征的精准度。

像汽油提纯这样有可用专家知识的实际问题还有很多,在面临此类问题时,也可以使用客观数据和主观知识相结合的方法进行更有效的关键特征选择。

2 融合人类知识的随机森林特征选择算法

本节提出了一种融合人类知识的特征选择方法,先利用随机森林对特征进行初选,然后借助模糊系统建模人类知识实现特征的进一步筛选。

2.1 随机森林特征初选

特征选择实质上是一个组合优化问题,是通过选取一组最优特征子集来达到特征约简的目的,即从维度为 S 的特征集中选出一组维度为 d ($d \ll S$) 的最优特征,有 $N = \binom{S}{d} = \frac{S!}{(S-d)! \cdot d!}$ 种可能的组合。实际应用中特征维度非常大,这样的计算量是难以接受的。因此,该文考虑基于随机森林算法^[11]和人类知识,提升优化效率,筛选出更优的特征子集。

在现有的大部分数据集中存在大量的样本信息,每个样本中又有大量描述样本的特征,这些特征之间具有高度非线性和强耦合的关系,特征之间相互制约,并且高维度的数据可能导致分类结果精确度不足,这使得在特征选择方法上,应该保证算法在适合处理非线性特征的基础上,又可以筛选出更具影响性的关键特征。传统的线性特征选择方法,如主成分分析、相关性分析,无法直接用于非线性特征变量的筛选,随机森林 (random forest, RF) 算法较其他特征选择算法,在过拟合问题影响相对较小,因此选择随机森林算法对特征进行初步选择。

随机森林算法如图 1 所示。

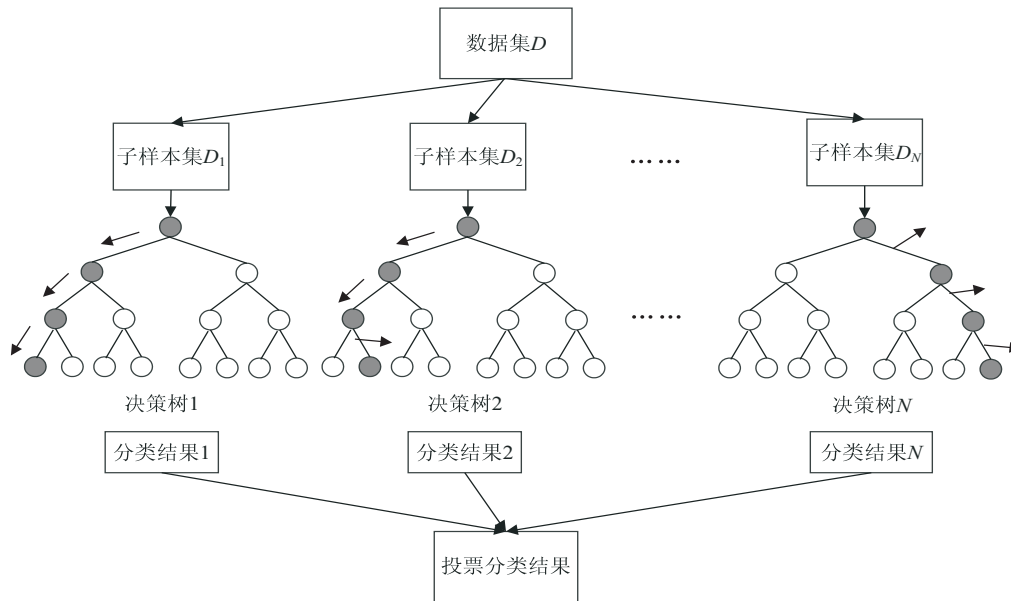


图1 随机森林算法

在数据集 D 中随机抽取 N 个子样本集(有放回随机抽样选择,迭代 N 次),每个子样本集对应一棵决策树,每棵决策树都是一个分类器,那么对于一个输入样本, N 棵树会有 N 个分类结果,根据每棵树的分类结果计算所有特征得分,并给定本棵决策树中最具影响的特征,即为本棵决策树的投票结果。通过随机森林集成所有决策树的投票结果,将投票次数最多的特征输出为关键特征。

具体来说,随机森林算法流程如下:

(1)在原始数据集 D 中,采取有放回的抽样方法随机选择 N 个样本集, N 个样本集中,每个样本集对应一棵决策树,由此构建 K 棵决策树,每次未被抽到的样本组成 K 个袋外数据(out-of-bag, OOB)以用于之后的无偏估计。

(2)每个样本中包含 M 个输入特征,每棵树的每个节点在 M 个输入特征中随机选取 m 个子特征($m \ll M$),计算每个子特征所含的信息量,在 m 个特征中选择一个最具分类能力的节点进行分裂。

(3)按照每棵树尽最大程度生长原则,对这棵树进行分枝生长,并且没有剪枝过程,直到这棵树可以按照制定标准分类数据集或所有属性都被使用过。根据分类结果,每棵决策树计算所有特征得分。

(4)将生成的多棵决策树集成随机森林,根据每棵树的投票结果(决策树最高得分特征)输出投票最高的特征即为关键特征。

(5)对于由多棵决策树构成的随机森林来说,就像一个黑盒子,无法控制模型内部的运行,只能在不同的参数和随机种子之间进行尝试,这会使很多相似的决策树掩盖了部分真实的结果,在特征选择上可能选择了部分非关键特征,对于小数据或者低维数据,容易

产生较差的分类效果。为了选择出更优的特征,该文融入人类专家知识对特征进一步筛选,从而尽量多地排除非关键特征。

2.2 融合人类知识的特征选择

2.2.1 人类知识建模

为了对经随机森林算法初选的特征进一步筛选,需要将人类专家知识融入模型。作为人类专家知识模型的代表,模糊系统(fuzzy system, FS)^[12]可以将输入、输出和状态变量定义在模糊集上,模糊系统抓住了人脑思维的模糊性特点,模仿人的综合推断来处理常规数学方法难以解决的模糊信息处理问题,较好地解决非线性问题。该文采用 Mamdani 型模糊系统^[13]作为关键特征筛选的人类知识模型。Mamdani 型模糊系统结构如图2所示,由模糊器、知识库、推理机和解模糊器四部分构成。

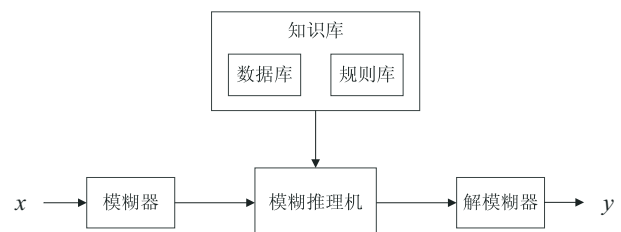


图2 模糊系统框架

首先模糊器将输入 x 模糊化成模糊集,然后推理机基于知识库对这些模糊集进行推理,最后解模糊器将推导出的模糊集转换为输出 y 。知识库是模糊系统的核心部分,主要包含数据库(data base, DB)和规则库(rule base, RB),其中 DB 包含语言规则中考虑的语言术语集和定义语言标签语义的隶属函数, RB 由 IF-THEN 形式的规则组成^[14]。组成 Mamdani 型模糊系统规则库的规则形式如下:

R^l : if x_1 is A_1^l and x_2 is A_2^l and \cdots and x_n is A_n^l , then y is B^l

其中, R^l 表示规则库中第 l 条规则 ($l = 1, 2, \cdots, M$), 其中 M 为模糊规则库规则数目; x_1 到 x_n 为输入语言变量, y 为输出语言变量; A_1^l 到 A_n^l 为规则前件的模糊集合, B^l 为规则后件的模糊集合。

在模糊系统中, 模糊规则是专家根据数据集中的人类知识搭建的推理方式, 它可以基于初步选择的特征变量, 并通过模糊系统推理给定初步选择特征中每个特征的最终得分。汽油提纯问题中, 其核心目的在于改变其他特征变量 (如氢油比、反应器上部温度、反应器底部压力等) 可使产品较少提高硫含量的同时更高降低辛烷值损失。以氢油比特征变量为例, 可以根据专家知识给出如下模糊规则以筛选关键特征变量:

如果氢油比是高, 辛烷值损失是高, 那么氢油比的影响因素高。

如果氢油比是低, 辛烷值损失是高, 那么氢油比的影响因素低。

如果氢油比是高, 辛烷值损失是低, 那么氢油比的影响因素低。

如果氢油比是低, 辛烷值损失是低, 那么氢油比的

影响因素高。

这里, 模糊系统的输出会给定此特征变量在汽油提纯问题中的影响效果得分。一般情况下, 模糊变量对应的隶属度函数由专家直接给出标准隶属度函数, 其论域取值为 0~1 之间。

同样的, 针对同时与产品硫含量和辛烷值损失有关的特征, 仍然可以搭建类似模糊系统评判特征变量的影响效果, 例如搭建模糊规则如下 (以氢油比、产品硫含量和辛烷值损失三个模糊变量中的一条规则为例):

如果氢油比是高, 产品硫含量是高, 辛烷值损失是高, 那么氢油比的影响因素高。

模糊系统可以利用大量人类知识进行建模, 从而使用人类知识指导特征变量的筛选, 接下来将描述如何搭建合适的模糊系统对特征进一步筛选。

2.2.2 融合人类知识的随机森林特征提取方法

为了更好地排除非关键属性带来的影响, 将人类知识引入随机森林特征选择中, 搭建一个良好的模糊系统, 有效地在特征初选之后进一步筛选出关键特征, 也就是将随机森林算法和模糊系统进行融合。融合人类知识的随机森林特征提取方法流程如图 3 所示。

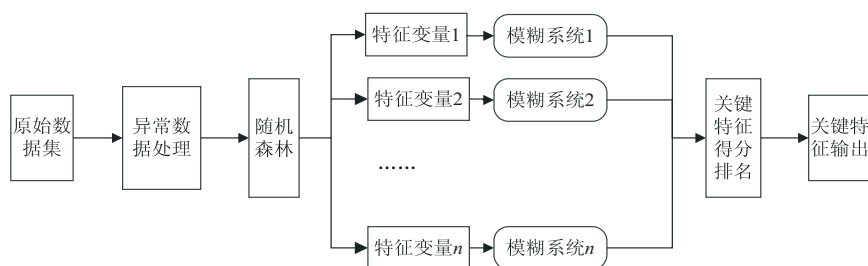


图 3 融合人类知识的随机森林特征提取方法流程

由于原始数据中大部分特征变量数据正常, 但数据集中少量数据可能出现数据不完整、数据缺失等问题, 需要对原始数据进行数据处理才能使用, 因此, 首先需要对原始数据进行预处理。面对特征数据过量冗余的原始特征, 随机森林算法中的 N 棵决策树对输入的原始特征进行投票, 经过 n 次迭代给出 n 个初步筛选后的特征。根据 n 个初步选择的特征变量分别搭建模糊系统, 将人类专家知识指导特征变量的进一步选择, 同样以汽油提纯问题中的氢油比特征变量的模糊规则为例:

如果氢油比是高, 辛烷值损失是高, 那么氢油比的影响因素高。

这里将氢油比特征在汽油提纯中的影响效果通过人类专家知识进行建模, 如果许多数据样本都体现出“氢油比高则辛烷值损失就高”的规律, 那么氢油比这个特征的影响因素就相应很高, 相反样本不能体现这个规则, 氢油比特征就没有那么重要。将数据样本输

入到所有模糊系统中就能确定每个特征的影响因素得分, 根据排名, 取其前 m 个特征变量作为最终的关键特征。

接下来设计实验验证融合人类知识的随机森林特征提取模型的有效性。

3 实验

3.1 数据集介绍及数据处理

原始数据采集来自于中石化高桥石化实时数据库 (霍尼韦尔 PHD) 及 LIMS 实验数据库。其中操作变量数据来自于实时数据库, 采集时间为 2017 年 4 月至 2020 年 5 月, 采集操作位点数共 354 个。2017 年 4 月至 2019 年 9 月, 数据采集频次为每 3 分钟 1 次; 2019 年 10 月至 2020 年 5 月, 数据采集频次为每 6 分钟 1 次。原料、产品和催化剂数据来自于 LIMS 实验数据库, 数据时间范围为 2017 年 4 月至 2020 年 5 月。其中原料及产品的辛烷值是重要的建模变量, 该数据采

集频次为每周 2 次。依据从催化裂化汽油精制装置采集的 325 个数据样本(每个数据样本都有 354 个特征变量),通过选择关键特征建立汽油辛烷值(RON)损失的预测模型,并根据模型预测验证该融合人类知识的随机森林特征提取方法的有效性。

原始数据中,大部分变量数据正常,但每套装置的数据均有部分位点存在问题:部分变量只含有部分时间段的数据,部分变量的数据全部为空值或部分数据为空值。因此对原始数据进行处理后才可以使⤵用。数据处理方法如下:

- (1)对于只含有部分时间点的位点,如果其残缺数据较多,无法补充,将此类位点删除;
- (2)删除 325 个样本中数据全部为空值的位点;
- (3)对于部分数据为空值的位点,空值处用其前后两个小时数据的平均值代替;
- (4)根据工艺要求与操作经验,总结出原始数据变量的操作范围,然后采用最大最小的限幅方法剔除一部分不在此范围的样本;
- (5)根据拉依达准则(3σ 准则)去除异常值。

3σ 准则:设对被测量变量进行等精度测量,得到 x_1, x_2, \dots, x_n , 算出其算术平均值 \bar{x} 及剩余误差 $v_i = x_i - \bar{x}$ ($i = 1, 2, \dots, n$), 并按贝塞尔公式算出标准误差 σ , 若某个测量值 x_b 的剩余误差 v_b ($1 \leq b \leq n$), 满足 $|v_b| = |x_b - \bar{x}| > 3\sigma$, 则认为 x_b 是含有较大误差值的坏值,应予剔除。贝塞尔公式如下:

$$\sigma = \left[\frac{1}{n-1} \sum_{i=1}^n v_i^2 \right]^{1/2} = \left\{ \left[\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 / n \right] / (n-1) \right\}^{1/2}$$

3.2 特征提取

实验中,在保证汽油产品脱硫效果(欧六和国六标准均为不大于 $10 \mu\text{g/g}$, 但为了给企业装置操作留有空间,要求产品硫含量不大于 $5 \mu\text{g/g}$)的前提下,尽量降低汽油辛烷值损失在 30% 以上,针对此问题搭建融合人类知识的随机森林二元回归模型。

基于训练好的随机森林特征选择方法算出每一个特征的重要性得分,并对这些特征进行排序,在所有特征中初选出重要性得分排名前 30 的特征。这里,将随机森林算法中子决策树个数设置为 81,最大特征数设置为 6。随机森林算法筛选的 30 个特征如下:

辛烷值 RON、精制汽油出装置硫含量、混氢点氢气流量、D-113 顶放空线流量、ME-115 过滤器压差、D-124 压力、ME-103 反吹气总管压力、烟气出辐射室温度、R-102 底喷头压差、D121 液面、D-124 液位、S_ZORB AT-0004、D104 液面、D-107 下部松动风流量、R102 转剂线压差、反吹氢气压力、R-102 床层吸附剂

料位密度、D-110 底部、D-107 底压力、加热炉氧含量、R-102 下部压力、氢油比、D-202 液位、P-101A 入口过滤器差压、精制汽油出装置温度、稳定塔顶回流量、D-201 含硫污水液位、芳烃、D-110 蒸汽盘管入口流量、D-201 含硫污水排量。

通过专家知识经验对变量进一步筛选是关键步骤,为了同时满足汽油辛烷值损失在 30% 以上,且产品硫含量不大于 $5 \mu\text{g/g}$,搭建如表 1 所示的模糊系统(以氢油比特征变量为例)。

表 1 满足辛烷值损失和硫含量条件下对关键变量提取的模糊系统

模糊系统输入		模糊系统输出		规则条数
模糊变量个数		模糊变量个数		
FS	氢油比	2	特征得分	2
	辛烷值损失	2		
	产品含硫量	2		

这里,由随机森林给定的 30 个特征变量,分别由专家给出对应的 30 个模糊系统,经过模糊系统给定 30 个特征变量最后得分,取其排名前 19 的关键特征变量,即最终关键特征变量,如表 2 所示。

表 2 融合人类知识的关键变量提取

特征变量名称		模糊系统评估得分
1	氢油比	0.93
2	反应器上部温度	0.89
3	反应器底部温度	0.84
4	反应器顶部压力	0.83
5	反应器顶底压差	0.82
6	反吹氢气压力	0.79
7	RT01 顶反应产物出口管温度	0.64
8	反应器质量空速	0.57
9	反应器料位	0.55
10	精制汽油出装置·硫含量	0.55
11	D-123 压力	0.46
12	D-113 顶放空线流量	0.46
13	ME-115 过滤器压差	0.46
14	E-101A 壳程出口管温度	0.44
15	1.0 MPa 蒸汽进装置温度	0.42
16	F-101 循环氢出口管温度	0.42
17	S_ZORB AT-0010	0.42
18	E-101C 管程出口管温度	0.42
19	R-102 床层吸附剂料位密度	0.40

3.3 不同预测模型对关键特征变量的验证

在模型预测过程中,分别选用训练好的随机森林回归^[15]、K 近邻回归^[16]和线性回归^[17]三种方法来验

证通过融合人类知识的随机森林算法在关键特征提取中的有效性。

为了能够通过降低模型在一次数据分割中性能表现上的方差来保证模型性能的稳定性,并且可以用于选择调节参数,比较模型性能差别,该文采用十折交叉验证方法将所有数据切分成 10 个子样本,每个子样本轮流作为测试集,其他 9 个样本作为训练集,重复 10 次,将 10 个结果进行平均最终得到一个单一的估计值,实验结果如表 3 所示。

表 3 实验结果

模型	未经过关键 特征选择 得分	随机森林 算法特征 选择得分	融合人类知识的 随机森林算法 特征选择得分
随机森林回归	0.59	0.64	0.77
K 近邻回归	-0.03	0.52	0.71
线性回归	0.34	0.49	0.58

由表 3 可以看出,在未经过特征选择任务中,随机森林回归算法、K 近邻回归算法和线性回归算法的表现得分偏低,K 近邻回归算法得分甚至出现负数,但随机森林回归算法较比其他两种回归算法表现较优,这说明随机森林回归算法在处理过度冗余的数据中有较好表现。经过随机森林初步选择之后的 30 个特征,在三种回归算法中的表现均有一定的提高,其中,K 近邻回归算法在初步筛选之后的特征中有较大的提升。将人类专家知识融入到随机森林特征选择后筛选出的 19 个关键特征中,在通过融合人类知识的随机森林算法提取的关键特征上的表现得分和经过随机森林初步特征选择的表现得分对比,可以看出三种回归算法性能有明显增强,且较未经过特征选择的原始特征上,三种回归算法表现得分显著提高。表明了人类专家知识在特征变量选择上的重要性,同时也证明了融合人类专家知识的随机森林算法在特征选择上较比单一的随机森林特征选择方法更优。

4 结束语

模式识别和数据挖掘中的一个重要问题是使用特征选择或特征提取进行降维,特别是在信息爆炸式增长的情况下,更需要降维处理。该文提出了一种新的特征选择方法,基于模糊系统建模,将人类专家知识整合到基于随机森林的特征选择方法中。通过随机森林算法对特征初步提取后,利用人类专家知识再对特征进一步筛选,从而得到关键特征。最后,通过与其他相关回归算法的比较,验证了该方法在真实数据集上具有更好的效果。通过这些研究和实验,证明了融合人类知识的随机森林特征选择方法在降维问题中的有

效性。

参考文献:

- [1] SONG X F,ZHANG Y,GONG D W,et al. Feature selection using bare-bones particle swarm optimization with mutual information[J]. Pattern Recognition,2020,112(4):107804.
- [2] GUPTA S,VERMA A K,AHMAD S. Feature selection for topological proximity prediction of single-cell transcriptomic profiles in drosophila embryo using genetic algorithm[J]. Genes,2020,12(1):28-46.
- [3] 李永豪,胡亮,张平,等. 基于动态图拉普拉斯的多标签特征选择[J]. 通信学报,2020,41(12):47-59.
- [4] 郭磊,王顺芳. 序列信息融合与两阶段特征选择的膜蛋白预测[J]. 计算机工程与应用,2019,55(6):145-150.
- [5] YUEF M,KUMAR A,BAKIR-GUNGOR B. Application of biological domain knowledge based feature selection on gene expression data[J]. Entropy,2020,23(1):2-19.
- [6] ZHANG R,ZHANG Y,LI X. Unsupervised feature selection via adaptive graph learning and constraint[J]. IEEE Transactions on Neural Networks and Learning Systems,2020,33(3):1355-1362.
- [7] 马语晗,赵辉. 基于特征选择加权支持向量机的运动模式识别[J]. 传感器世界,2018,24(9):28-33.
- [8] 陈莉芬. 面向多源特征的模式识别算法及应用研究[D]. 济南:山东大学,2020.
- [9] 熊熙,乔少杰,韩楠,等. 一种基于模糊选项关系的关键属性提取方法[J]. 计算机学报,2019,42(1):190-202.
- [10] 孙广路,何勇军,刘广明. 基于最大信息系数的特征选择,分类方法及其装置:CN,104050242B[P]. 2018.
- [11] BREIMAN L. Random forests[J]. Machine Learning,2001,45:5-32.
- [12] ZADEH L A. Fuzzy sets[J]. Information and Control,1965,8(3):338-353.
- [13] WU D,LIN C,HUANG J,et al. On the functional equivalence of TSK fuzzy systems to neural networks, mixture of experts,CART and stacking ensemble regression[J]. IEEE Transactions on Fuzzy Systems,2020,28(10):2570-2580.
- [14] HERRERA F. Genetic fuzzy systems: taxonomy, current research trends and prospects[J]. Evolutionary Intelligence,2008,1(1):27-46.
- [15] IVERSON L R,PRASAD A M,MATTHEWS S N,et al. Estimating potential habitat for 134 eastern US tree species under six climate scenarios[J]. Forest Ecology and Management,2008,254(3):390-406.
- [16] ATHEY S,IMBENS G W. Machine learning methods economists should know about[J],Research Papers,2019,11(1):685-725.
- [17] 薛素静,上官同英. 多元线性回归算法的研究和应用[J]. 水利电力机械,2007,29(5):59-60.