doi: 10.15940/j.cnki.0001-5245.2020.02.010

基于Stacking集成学习的恒星/星系分类研究*

李 超[1,2] 张文辉[3,4†] 李 然[1,5] 王俊义[1,2‡] 林基明[1,2,6]

(1 桂林电子科技大学信息与通信工程学院 桂林 541004)
(2 桂林电子科技大学认知无线电与信息处理教育部重点实验室 桂林 541004)
(3 桂林电子科技大学广西云计算与大数据协同创新中心 桂林 541004)
(4 桂林电子科技大学广西高校云计算与复杂系统重点实验室 桂林 541004)
(5 桂林电子科技大学广西无线宽带通信与信号处理重点实验室 桂林 541004)
(6 广西高校卫星导航与位置感知重点实验室 桂林 541004)

摘要 机器学习在当今诸多领域已经取得了巨大的成功, 但是机器学习的预测效果往往依赖于具体问题. 集成学习通过综合多个基分类器来预测结果, 因此, 其适应各种场景的能力较强, 分类准确率较高. 基于斯隆数字巡天(Sloan Digital Sky Survey, SDSS)计划恒星/星系中最暗源星等集分类正确率低的问题, 提出一种基于Stacking集成学习的恒星/星系分类算法. 从SDSS-DR7 (SDSS Data Release 7)中获取完整的测光数据集, 并根据星等值划分为亮源星等集、暗源星等集和最暗源星等集. 仅针对分类较为复杂且困难的最暗源星等集展开分类研究. 首先, 对最暗源星等集使用10折嵌套交叉验证, 然后使用支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest, RF)、XGBoost (eXtreme Gradient Boosting)等算法建立基分类器模型; 使用梯度提升树(Gradient Boosting Decision Tree, GBDT)作为元分类器模型. 最后, 使用基于星系的分类正确率等指标, 与功能树(Function Tree, FT)、SVM、RF、GBDT、XGBoost、堆叠降噪自编码(Stacked Denoising AutoEncoders, SDAE)、深度置信网络(Deep Belief Network, DBN)、深度感知决策树(Deep Perception Decision Tree, DPDT)等模型进行分类结果对比分析. 实验结果表明, Stacking集成学习模型在最暗源星等集分类中要比FT算法的星系分类正确率提高了将近10%. 同其他传统的机器学习算法、较强的提升算法、深度学习算法相比, Stacking集成学习模型也有较大的提升.

关键词 恒星: 基本参数, 星系: 基本参数, 技术: 测光, 方法: 数据分析

中图分类号: P152; **文献标识码**: A

2019-12-13收到原稿, 2020-01-10收到修改稿

*国家自然科学基金项目(61966007)、认知无线电与信息处理教育部重点实验室开发基金项目(CRKL180201)、广西无线宽带通信与信号处理重点实验室主任基金项目(GXKL06180107)、广西云计算与大数据协同创新中心、广西高校云计算与复杂系统重点实验室项目(1716)资助

†zhangwh@guet.edu.cn

‡wangjy@guet.edu.cn

1 引言

近年来, 随着各国空间科学技术和大型巡天项目的不断开展, 天文学显然已经发展到了一个前所未有的阶段, 即大数据-巨信息量-全波段时代[1]. 面对如此庞大而又复杂的天文数据, 如何进行高效而且准确的数据分析显得极为重要. 恒星/星系分类一直是天文数据分析的基本内容之一, 而且人们对它的研究最早可以追溯到18世纪[2]. 之前被广泛应用于解决恒星/星系分类问题的是基于形态、启发式分割等原始方法. 近些年来, 随着原始方法在解决恒星/星系分类问题上速度慢、分类准确率低等缺点的突显, 基于机器学习和深度学习等优秀的模型和算法的研究也随之展开. 如文献[3]在SDSS-DR6 (Sloan Digital Sky Survey Data Release 6)的测光数据上, 使用自动聚类的方法, 进行恒星/星系的分类, 结果表明, 自动聚类算法具有较高的效率; 文献[4]在SDSS-DR7的测光数据集上, 对13种不同的决策树算法进行了恒星/星系分类效果的对比, 结果表明, 功能树(Function Tree, FT)算法在恒星/星系分类的问题上要优于其他决策树算法; 文献[5]探讨了深度置信网络(Deep Belief Network, DBN)、神经网络(Neural Networks, NN)和支持向量机(Support Vector Machine, SVM)等算法在Sloan天文数据中光谱分类的应用, 结果表明, 以上3种自动光谱分类算法有很大的利用价值; 文献[6–7]通过在SDSS-DR7数据上, 使用了堆叠降噪自编码(Stacked Denoising Autoencoders, SDAE)算法, 对于解决恒星/星系的最暗源星等集分类问题提供了一种有效的思路; 文献[8]提出了一种基于深度感知决策树(Deep Perception Decision Tree, DPDT)的算法, 明显提高了SDSS-DR7最暗源星等集恒星/星系的分类准确率; 文献[9]在集成学习的背景下, 探索了随机森林(Random Forest, RF)、Adaboost (Adaptive boosting)、极端随机树(Extremely randomized trees, ET)、梯度提升树(Gradient Boosting Decision Tree, GBDT)等几种树模型在天文学中恒星/星系分类中的应用, 并且给出了合理的解释. 在天文学领域, 已经研究并使用了很多优秀的算法, 但是这些算法都存在一些问题, 如模型单一、使用场景有限、泛化能力弱等. 在SDSS-DR7数据中最暗源星等集分类正确率低的问题始终无法得到有效的解决. 因此本文构建了一种基于Stacking的恒星/星系分类两层集成算法框架, 并创新性地将Stacking框架应用到SDSS-DR7测光数据中, 较好地解决了SDSS-DR7测光数据中最暗源星等集的恒星/星系分类准确率低的问题. 因此, 基于多模型融合的Stacking集成学习方法对于天文学研究有很高的应用价值.

2 算法理论

2.1 Stacking集成学习算法

Stacking集成学习[10]是一种异质集成的策略. 异质集成是通过集成若干个不同类型的基分类器, 组合成一个强分类器, 以此来提升强分类器的泛化能力. Stacking集成学习算法采用两层框架的结构, 如图1所示. 其训练过程如下: 首先分别对多个基分类器进行训练; 然后将多个基分类器的预测结果作为元分类器的输入, 再次进行训练. 最终的集成算法会兼顾基分类器和元分类器的学习能力, 使得分类精度和准确率得到明显提升. Stacking集成学习算法的效果好坏取决于两个方面: 一个是基分类器的预测效果, 通常基分类器的预测效果越好, 集成学习模型的预测效果越好; 另一个是基分类器之间需要

有一定的差异性, 因为每个模型的主要关注点不同, 这样集成才能使每个基学习器充分发挥其优点. 试想, 如果基分类器的差异性较低, 那么每个基分类器的预测结果就会很相似, 那么这样集成和单个分类器的预测基本没有区别, 只会徒增模型的复杂度.

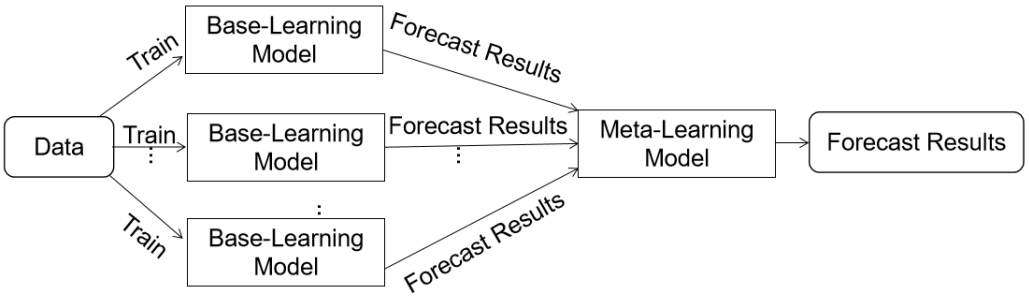

图 1　基于Stacking的集成学习算法

Fig. 1　An ensemble learning algorithm based on Stacking

2.2　支持向量机算法

支持向量机(Support Vector Machine, SVM)是一种二类分类模型, 其基本模型是定义在特征空间上的间隔最大的线性分类器. 线性可分SVM算法旨在找到一个可以完全划分所有数据的超平面, 使得数据集中所有数据距离此超平面最远, 即硬间隔(hard margin) SVM. 当训练数据近似线性可分时, SVM通过软间隔(soft margin)最大化也可以学习到一个线性分类器, 也称软间隔SVM. 随着数据复杂程度的提高, 当训练数据线性不可分时, 通过引入软间隔最大化和核技巧, 学习到一个分类器, 即非线性SVM. 非线性SVM可以将在原始特征空间中线性不可分的训练样本映射到一个高维的特征空间中, 从而使得映射后的训练样本在高维特征空间中线性可分. 本文使用的SVM算法采用的是径向基函数(Radial Basis Function, RBF), 也称高斯核函数:

$$Z(\boldsymbol{u},\boldsymbol{v}) = \exp\left(\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}\right), \tag{1}$$

其中$\boldsymbol{u}$和$\boldsymbol{v}$表示为两个样本向量, Z表示为RBF核函数的值, σ是一个自由参数.

2.3　随机森林算法

随机森林(Random Forest, RF)是集成学习中Bagging思想的一种算法策略. Bagging思想是对训练集进行随机采样, 产生出多个不同的训练子集, 再对每个训练子集训练出一个基分类器, 预测结果通过多个基分类器取平均或者投票得出. 这时的预测模型有望获得较好的预测结果和较强的泛化能力. 随机森林是在将决策树作为基学习器构建Bagging集成算法的同时, 还引入了特征的随机采样, 进一步提升了模型的抗噪声能力, 有效地防止了过拟合的发生.

2.4 梯度提升树算法

梯度提升决策树(Gradient Boosting Decision Tree, GBDT)[11]是集成学习Boosting思想中的一种算法, 它同样将决策树作为基函数. GBDT算法的核心在于每颗树学习的是之前所有树结论和的残差. 传统的提升树算法采用平方损失函数, 可以直接计算残差, 但是缺点是仅能解决回归问题. GBDT算法对其做了改进, 它每次在建立单个弱分类器时, 是在之前建立模型的损失函数的梯度下降方向(或称负梯度值)来近似残差. 因此, 多种损失函数的选取, 不仅可以帮助GBDT算法有效地解决回归问题, 同时也可以解决分类问题. GBDT算法的学习能力较强, 是如今机器学习领域非常重要的一个算法.

2.5 XGBoost算法

XGBoost (eXtreme Gradient Boosting)[12]是一种对GBDT做了改进的提升算法, 在优化时同时使用一阶导数信息和二阶导数信息. 其模型如下所示:

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F\,, \tag{2}$$

其中$\widehat{y}_i$表示第i个样本的预测结果, K表示决策树的数目, F代表了决策树的集合空间, $F = \{f(x_i) = w_{q(x_i)}\}$, f表示树模型, $w_{q(x_i)}$把每一个节点映射成一个值, 即$f(x_i)$的值. x_i表示第i个样本, f_k表示第k棵树的模型, 并且每一个f_k对应着一个独立的树结构和叶子节点的权值. q表示一个独立的树结构, 其作用是将样本实例映射到相应的叶子节点. XGBoost的目标函数如下:

$$\begin{cases} L = \sum\limits_{i=1}^{n} l(\widehat{y}_i, y_i) + \sum\limits_{k} \Omega(f_k) \\ \Omega(f_k) = \gamma T + \dfrac{1}{2}\lambda \sum\limits_{j=1}^{T} {w_j}^2 \end{cases}, \tag{3}$$

其中n表示样本数目, l是一个用来计算预测值$\widehat{y}_i$和真实值y_i之间训练误差的函数, Ω表示为树模型复杂度, γ表示为叶子数目的正则化参数, 用来抑制节点继续向下分裂, T表示一棵决策树中叶子节点的总数量, λ表示为叶子权重的正则化参数, w_j表示叶子节点j的权重. 算法的目标是最小化损失函数:

$$L^{(t)} = \sum_{i=1}^{n} l\left[\widehat{y}_i^{(t-1)} + f_t(x_i), y_i\right] + \Omega(f_t)\,, \tag{4}$$

其中$L^{(t)}$表示为第t轮的目标函数, $\widehat{y}_i^{(t-1)}$表示为前$t-1$棵树的输出值之和, 构成前$t-1$棵树的预测值, f_t表示为第t棵树的模型, $f_t(x_i)$表示为第t棵树的输出结果, $\widehat{y}_i^{(t-1)} + f_t(x_i)$相加构成最新的预测值.

定义g_i和h_i:

$$g_i = \frac{\partial l\left(y_i, \widehat{y}_i^{(t-1)}\right)}{\partial \widehat{y}_i^{(t-1)}}\,, \tag{5}$$

$$h_i = \frac{\partial^2 l\left(y_i, \widehat{y}_i^{(t-1)}\right)}{\partial\left(\widehat{y}_i^{(t-1)}\right)^2}\,, \tag{6}$$

将损失函数在$\widehat{y}_i^{(t-1)}$处利用泰勒公式展开:

$$L^{(t)} \approx \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)\,. \tag{7}$$

去掉常数项, 第t次迭代后的损失函数变为:

$$L^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)\,. \tag{8}$$

定义$I_j = \{i \mid q(x_i) = j\}$作为叶子节点$j$的实例集, 其中$I$表示节点划分前的实例集, 根据(8)式得:

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i {f_t}^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum\nolimits_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum\nolimits_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T\,. \end{aligned} \tag{9}$$

对于固定了的决策树的结构$q(x_i)$, 可以计算得出叶子节点j的最优权重w_j^*:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}\,, \tag{10}$$

将w_j^*回代入目标函数得:

$$L^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T\,. \tag{11}$$

(11)式作为衡量树结构质量的指标, 可以用来计算树结构q的得分. 同时, 需要使用贪心算法迭代地在每一个已有的叶子节点添加分支. 假定$I_{\rm L}$和$I_{\rm R}$是划分后左右子树叶子节点的集合, 即$I = I_{\rm L} \cup I_{\rm R}$, 则划分后的损失函数如下:

$$L_{\rm split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_{\rm L}} g_i\right)^2}{\sum_{i\in I_{\rm L}} h_i + \lambda} + \frac{\left(\sum_{i\in I_{\rm R}} g_i\right)^2}{\sum_{i\in I_{\rm R}} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma\,. \tag{12}$$

3　基于Stacking集成学习算法的训练

本文充分考虑了决定Stacking集成学习模型效果好坏的两个方面: 一是选择学习能力较强的基学习器; 二是充分考虑基学习器之间的差异性. SVM在解决非线性的中小规模数据集的分类和回归中具有非常好的效果. RF和XGBoost分别是集成学习Bagging和Boosting中泛化能力和学习能力较强的算法. 3种算法不仅有充分的理论

支撑, 而且在科学研究中正扮演着重要的角色. 第2层元学习器同样选择学习能力较强的GBDT算法, 用于对第1层基学习器的集成, 并且使用10×10折嵌套交叉验证划分数据的方式防止过拟合的发生. 综上所述, 本文基于Stacking集成学习的分类模型第1层基学习器选择SVM、RF、XGBoost, 第2层元学习器选择GBDT, 模型结构如图2所示.

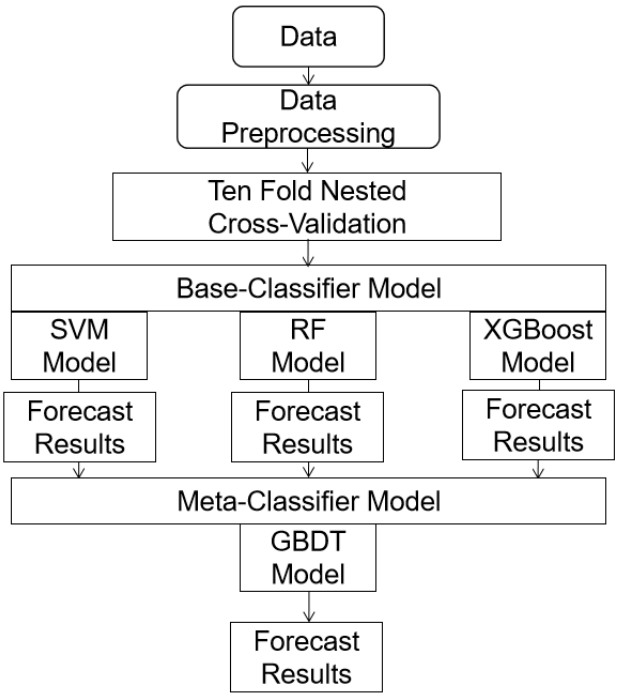

图 2 基于Stacking集成学习的恒星/星系分类模型

Fig. 2 A star/galaxy classification model based on the Stacking ensemble learning

传统的10折交叉验证就是将原始数据划分为10等分, 轮流将其中的9份作为训练集, 剩下1份作为测试集. 本文采用10×10折嵌套交叉验证的划分方法, 即在每一个训练集的内部再做一次10折交叉验证.

基于Stacking集成学习框架的训练流程如下:

(1)对原始数据进行预处理并且按照10×10折嵌套交叉验证的方式进行划分;

(2)使用划分后的数据集分别对第1层基学习器中的SVM、RF、XGBoost 3种算法进行训练, 并得到预测结果;

(3)将第1层基学习器的预测结果拼接起来作为第2层元学习器GBDT的输入, 再次进行训练, 并得到最终的预测结果.

4 实验结果与分析

4.1 数据集介绍

完整的SDSS-DR7测光数据集见http://skyserver.sdss.org/dr7/en/, 根据星等值(modelMag)大小可以划分为: 亮源星等集(14–19)、暗源星等集(19–21)、最暗源星等集(20.5–21). 与SDSS-DR7恒星/星系亮源和暗源星等集数据相比, 最暗源星等集数据量规模较小, 数据测量困难, 分类准确率较低. 因此本文采用的是SDSS-DR7恒星/星系最暗源星等数据集, 可直接使用简单的SQL (Structured Query Language)语句从Skysever平台获取, 并且与文献[4]特征参数保持一致. 数据特征参数如表1所示.

表 1 用于SDSS-DR7恒星/星系分类的特征参数
Table 1 The feature parameters for SDSS-DR7 star/galaxy classification

Variable	Attribute
psfMag	PSF (point-spread function) magnitude
fiberMag	Fiber magnitude
petroMag	Petrosian magnitude
modelMag	Model magnitude
petroRad	Petrosian radius
petroR50	Radius carrying 50% of Petrosian flux
petroR90	Radius carrying 90% of Petrosian flux
lnLStar	Likelihood PSF
lnLExp	Likelihood exponential
lnLDeV	Likelihood deVaucouleurs
mRrCc, mE1, mE2	Adaptive moments
specClass	Spectroscopic classification

4.2 参数设置

基于Stacking集成学习模型通过将SVM、RF、XGBoost算法作为基学习器训练, 得到预测结果, 作为元学习器GBDT的输入, 再次进行训练, 得到最终预测结果. 各个算法的主要参数设置如下: SVM算法模型采用RBF, gamma参数设置为1; RF算法模型采用计算属性的基尼指数来选择分裂节点, 决策树的个数和深度分别为65和7; XGBoost算法模型的弱学习器数目设置为710, 学习速率设置为0.01, 树的深度设置为6; GBDT算法模型的弱学习器数目设置为200, 学习速率设置为0.04, 树的深度设置为3.

4.3 实验方法及模型对比

为了能更好地评估基于Stacking集成学习模型在恒星/星系最暗源星等集分类上的性能, 本文对比了FT、SVM、RF、GBDT、XGBoost[13−14]、DBN、SDAE、DPDT等算法, 详细的对比实验结果如表2. 同样, 为了保证对比分类结果的有效性, 采用了与文献[4]一致的分类性能衡量指标(CP), 即星系的分类正确率. 其定义如(13)式所示:

$$\mathrm{CP}(m) = 100 \times \frac{N_{\mathrm{gal-gal}}(m)\delta m}{N^{\mathrm{tot}}_{\mathrm{galaxy}}(m)\delta m}\,, \tag{13}$$

其中, $N_{\mathrm{gal-gal}}(m)\delta m$代表星等值在$(m-\frac{\delta m}{2}, m+\frac{\delta m}{2})$区间内的数据样本中被正确分类为星系的数量, $N^{\mathrm{tot}}_{\mathrm{galaxy}}(m)\delta m$代表星等值在$(m-\frac{\delta m}{2}, m+\frac{\delta m}{2})$区间内数据样本中星系的总数量. 本文仅使用modelMag在20.5–21之间的最暗源星等集.

表 2 SDSS-DR7星系分类正确率

Table 2 The accuracy of SDSS-DR7 galaxy classification

Method \ Set	CP(20.5–21)/%
FT	74.04
DBN	74.45
SDAE	73.08
DPDT	77.47
SVM	71.15
RF	76.52
GBDT	78.30
XGBoost	80.75
Stacking	84.42

通过仿真实验得出的表2可以看出, 对最暗源星等集, 基于Stacking集成学习模型的星系分类准确率要远优于FT, 提高了约10%的准确率. 而与之前已经研究过的SDAE、DPDT模型相比, 准确率提高了约7%–10%. 与其他较为先进的DBN、SVM、RF、GBDT、XGBoost等算法相比, 也提高了约4%–13%的星系分类准确率. 由此可见, 基于Stacking集成学习模型综合了各个基分类器的优点后, 并充分发挥了集成模型的性能, 因此具有更强的泛化能力和更好的预测效果.

5 结论

本文通过使用SDSS-DR7测光数据集, 并且采用10 × 10折嵌套交叉验证的方法, 研究了基于Stacking集成学习算法的恒星/星系的分类问题. 最后通过对基分类器和元分类器参数调优, 基于星系分类准确率的评价指标, 与FT、SVM、RF、GBDT、XGBoost、DBN、SDAE、DPDT等模型进行对比. 实验结果表明, 基于Stacking集成学习模型在恒星/星系最暗源星等集上的分类效果要远好于其他模型. 因此, 该Stacking集成学习模型在天文学有非常高的应用价值.

在下一步工作中, 将探讨解决Stacking集成学习模型的算法复杂度问题. 在中小规模数据集上, 该集成模型应用较好. 但是, 遇到大规模或者超大规模数据集, 势必会大大增加集成模型的训练时间. 因此, 在未来的研究中, 会尝试使用分布式的方法, 对基学习器并行训练, 这样不仅会使集成模型达到较高的精确度, 而且也会使得集成模型训练起来有较高的效率.

参 考 文 献

[1] 张彦霞, 赵永恒. 科研信息化技术与应用, 2011, 2: 13
[2] Messier C. Connoissance des Temps for 1784, 1781: 227
[3] 严太生, 张彦霞, 赵永恒, 等. 中国科学G辑: 物理学力学天文学, 2009, 39: 1794
[4] Vasconcellos E C, De Carvalho R R, Gal R R, et al. AJ, 2011, 141: 189
[5] 李俊峰, 汪月乐, 胡升, 等. 光谱学与光谱分析, 2016, 36: 3261
[6] 秦浩然, 林基明, 王俊义. 天文学报, 2016, 57: 344
[7] Qin H R, Lin J M, Wang J Y. ChA&A, 2017, 41: 282
[8] 黄智昌, 王俊义, 郑霖, 等. 计算机应用研究, 2017, 34: 765
[9] Morice-Atkinson X, Hoyle B, Bacon D. MNRAS, 2018, 481: 4194
[10] Zhou Z H. Ensemble Methods: Foundations and Algorithms. Boca Raton: CRC Press, 2012
[11] Friedman J H. The Annals of Statistics, 2001, 29: 1189
[12] Chen T Q, Guestrin C. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. California: ACM, 2016: 785
[13] 李超, 张文辉, 林基明. 天文学报, 2019, 60: 71
[14] Li C, Zhang W H, Lin J M. ChA&A, 2019, 43: 539

Research on Star/Galaxy Classification Based on Stacking Ensemble Learning

LI Chao[1,2] ZHANG Wen-hui[3,4] LI Ran[1,5] WANG Jun-yi[1,2] LIN Ji-ming[1,2,6]

(1 College of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin 541004)
(2 Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education, Guilin University of Electronic Technology, Guilin 541004)
(3 Guangxi Cooperative Innovation Center of Cloud Computing and Big Data, Guilin University of Electronic Technology, Guilin 541004)
(4 Guangxi Colleges and Universities Key Laboratory of Cloud Computing and Complex Systems, Guilin University of Electronic Technology, Guilin 541004)
(5 Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin University of Electronic Technology, Guilin 541004)
(6 Guangxi Colleges and Universities Key Laboratory of Satellite Navigation and Position Sensing, Guilin 541004)

Abstract Machine learning has achieved great success in many areas today, but the predictive effect of machine learning often depends on the specific problem. An ensemble learning predicts results by integrating multiple base classifiers. Therefore, its ability to adapt to various scenarios is strong, and the classification accuracy is high. In response to the low classification accuracy of darkest source magnitude sets in star/galaxy in the Sloan Digital Sky Survey (SDSS), a star/galaxy classification algorithm based on the Stacking ensemble learning is proposed in this paper. The complete photometric data set is obtained from SDSS-Data Release (DR) 7 and divided into bright source magnitude set, dark source magnitude set, and darkest source magnitude set according to the magnitude. Firstly, the ten-fold nested cross-validation method is used for the darkest source magnitude set, and then the Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost) algorithms are used to establish the base-classifier model; the Gradient Boosting Decision Tree (GBDT) is used as the meta-classifier model. Finally, based on galaxies' classification accuracy and other indicators, the classification results are compared with the models of Function Tree (FT), SVM, RF, GBDT, Stacked Denoising Autoencoders (SDAE), Deep Belief Nets (DBN), and Deep Perception Decision Tree (DPDT) etc., and then analyzed. The experimental results show that, the Stacking ensemble learning model improves the classification accuracy of galaxies in the darkest source classification by nearly 10% compared to the function tree algorithm. Compared with other traditional machine learning algorithms, strong lifting algorithms and deep learning algorithms, the Stacking ensemble learning model also has different degrees of improvement.

Key words stars: fundamental parameters, galaxies: fundamental parameters, techniques: photometric, methods: data analysis