doi:10.6041/j.issn.1000-1298.2019.11.016

基于 Stacking 集成学习的水稻表型组学实体分类研究

袁培森[1]　杨承林[1]　宋玉红[1]　翟肇裕[2]　徐焕良[1]

(1.南京农业大学信息科学技术学院,南京 210095;2.马德里理工大学技术工程和电信系统高级学院,马德里 28040)

摘要:为研究整合水稻表型组学相关知识,系统地建立水稻表型组学知识图谱,通过分布式爬虫框架从国家水稻数据中心网站获取水稻表型组学数据集,并以互动百科为辅助数据源获取水稻表型组学数据。对水稻表型组学数据采用 TF-IDF 技术结合潜在语义模型进行预处理,并对水稻表型组学实体进行人工分类和标注。为实现水稻表型组学实体分类,研究了基于堆叠式两阶段集成学习的分类器组合模型,结合 K-近邻算法、支持向量机、随机森林、梯度提升决策树机器学习方法,提升水稻表型组学实体数据分类的性能。研究表明,基于堆叠式两阶段集成学习的分类器组合模型对不同类别的水稻表型组学数据都具有较好的多分类能力,对于不平衡的水稻表型组学数据集,本文方法的分类器组合模型对水稻表型组学数据分类效果最佳,Gene 类别的 F_1 为 90.47%,总体准确率达 80.55%,比支持向量机、K-近邻、随机森林和梯度提升决策树 4 种基分类器的分类准确率平均高 6.78 个百分点。

关键词:水稻表型组学;实体分类;堆叠式集成学习;知识图谱;潜在语义模型

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1000-1298(2019)11-0144-09

Classification of Rice Phenomics Entities Based on Stacking Ensemble Learning

YUAN Peisen[1]　YANG Chenglin[1]　SONG Yuhong[1]　ZHAI Zhaoyu[2]　XU Huanliang[1]

(1. *College of Information Science and Technology*, *Nanjing Agricultural University*, *Nanjing* 210095, *China*
2. *Superior School of Technical Engineering and Telecommunication Systems*, *Technical University of Madrid*, *Madrid* 28040, *Spain*)

Abstract: With the development of rice phenomics research, it is of great significance for comprehensively analyzing, mining and applying the rice phenomics data. In order to integrate the knowledge related to rice phenomics and explore the factors affecting rice phenotypic traits, the rice phenomics knowledge graph system was implemented. Rice phenomics knowledge graph system consisted of functional modules such as entity recognition, entity query, relational query and knowledge visualization. The rice phenomics data were downloaded by a distributed data website crawler from the National Rice Data Center website, and the interactive encyclopedia website was taken as auxiliary data sources to obtain rice phenomics dataset. The dataset was preprocessed with TF-IDF and latent semantic indexing method and classified and labeling manually firstly, and then machine learning approaches were applied for training and testing. The rice phenomics entity classification was studied based on stacking ensemble learning integrated with basic learning classifier, such as K-nearest neighbor, support vector machine, random forests and gradient boosting decision tree. Based on stacking ensemble learning classifier, different types of rice phenomics data showed fine ability for entity classification. For the unbalanced rice phenomics entities, comparing with the support vector machine algorithm, the K-nearest neighbor algorithm, the random forest algorithm and the gradient boosting decision tree algorithm, the proposed method showed the best performance, i.e. the F1-Measure of Gene entities can reach 90.47%. The overall accuracy was 80.55%, and it was 6.78 percentage points higher than those of the other four basic classifiers.

Key words: rice phenomics; entities classification; stacking ensemble learning; knowledge graph; latent semantic indexing

收稿日期:2019-07-11　修回日期:2019-08-24

基金项目:国家自然科学基金项目(61502236)、中央高校基本科研业务费专项资金项目(KJQN201651)和大学生创新创业训练专项计划项目(S20190025)

作者简介:袁培森(1980—),男,讲师,博士,主要从事智能信息、海量数据处理与分析研究,E-mail:peiseny@njau.edu.cn

通信作者:徐焕良(1963—),男,教授,博士,主要从事农业信息化与大数据技术研究,E-mail:huanliangxu@njau.edu.cn

0　引言

中国是世界第二大水稻种植国家，水稻总产量占全球 30% 以上[1]。培育优良的水稻品种、提高水稻产量成为我国重要的战略目标。植物表型是植物在一定环境下表现的可观察形态特征，在植物保护、育种等领域具有重要的应用价值，其研究涉及植物学、数据科学、机器学习等领域[2]。其中，水稻表型组学的研究发展迅速。水稻的表型组学文献数量占第 2 位，进而出现了大量的水稻表型组学相关的知识[3]，为水稻表型组学研究提供大量的数据支持。目前，水稻表型组学研究正在成为水稻基因组功能分析、分子育种和农业应用中的重要技术支撑[4-7]，其知识图谱的研究非常迫切。

知识图谱技术可将农业领域中的水稻表型组学相关信息表达为更贴近人类认知的形式，提供一种更好地组织、管理和理解水稻表型组学海量信息的能力[8]。近年来，国外的知识图谱研究比较注重对一些社会热点进行分析，而国内利用知识图谱处理农业信息化数据从 2012 年开始[9-11]。国内外学者已在农业领域知识图谱的构建方面进行了相关研究，但对于水稻表型组学知识图谱的构建和分析甚少。

知识图谱的关键技术包括知识抽取、知识表示、知识融合，以及知识推理技术。实体是知识图谱中最基本的元素，其抽取和分类的准确率、召回率等将直接影响到后续知识库的品质和知识获取效率[12]。传统的分类器学习系统使用已有的学习器（如支持向量机）对给定的训练样本集进行训练产生一个模型，用产生的模型预测新的样例后，再根据模型的预测结果对其分类性能进行分析[13]。由于现实中数据量的不断增加和数据的多元化，尤其是水稻表型组学数据的复杂性和专业性，使传统的分类算法已经无法满足现有数据的处理以及实际问题的解决需求。对单一分类器的组合学习模型在分类问题上显示出了极大的优势，使得组合分类模型在分类问题中得到更多的关注。现最流行的组合分类算法有 Boosting[14] 和 Bagging[15] 方法。对于不稳定性的分类算法，Bagging 方法分类效果较好（如决策树[16]），但对于稳定的分类器集成效果不是很理想。Boosting 方法的基分类器训练集取决前一个基分类器，其分错的样例按较大的概率出现在下一个基分类器的训练集中，虽然提高了组合分类算法的泛化性能，但会过分偏向于一些很难分的样例，从而导致算法性能的降低。

针对 Bagging 和 Boosting 两种组合算法的分类器集成效果并不理想，且改善效果均十分有限的问题，本文利用基于 Stacking[17] 集成学习策略的两层式叠加框架，提出一种基于 Stacking 集成学习的分类器组合模型。通过爬虫框架获取国家水稻数据中心的水稻表型组学数据，利用 TF－IDF 技术和 LSI 模型结合的方法对水稻表型组学数据进行预处理，消除冗余特征，然后在基于 Stacking 集成学习算法的多分类器组合方法上，将支持向量机（SVM）、K-近邻（K-NN）、梯度提升决策树（GBDT）和随机森林（RF）4 种学习算法作为初级分类器进行组合学习，进而采用随机森林算法作为次级分类器提升分类效果。

1　水稻表型组学知识库的构建

1.1　水稻表型组学数据获取

水稻表型组学数据主要从“国家水稻数据中心”网站（http://www.ricedata.cn/）获取，并以互动百科网站作为辅助数据源。Scrapy[18] 爬虫框架作为本文数据获取的工具，其充分利用处理器资源，快速抓取结构化数据；支持分布式爬虫，适合大规模水稻表型组学数据的获取。

1.2　水稻表型组学知识图谱示例

存储水稻表型组学数据的数据库使用专用于图存储的图数据库 Neo4j[19]，图数据库利用结点和边的方式表示实体、属性与关系，使得数据表示更为直观清晰，适用于存储水稻表型组学实体、属性及关系。本文使用 RDF 三元组的数据表示形式，存放实体-属性或实体-关系，共存储 4 853 个数据结点，5 438 条关系，每个结点包括 ID、title、detail 等属性。Neo4j 图数据库使用的专门图形查询语言 Cypher，简单快捷，便于操作，且能够以图、表、文本的形式展示数据。实例如图 1 所示。

1.3　水稻表型组学实体识别

水稻表型组学实体识别是指识别水稻表型组学文本数据中具有特定实际意义的水稻表型组学实体，包括表型、基因、蛋白质、环境等专有名词。本文从国家水稻数据中心获取的大量非结构化水稻表型组学数据中抽取实体，针对收集的语料库进行分词及词性标注，根据词性标注结果基于一定的启发式规则进行词语组合，从文本中提取具有实际意义的实体。图 2 所示为水稻表型组学实体识别过程。

为实现水稻表型组学实体识别，本文借助 THULAC 分词工具进行分词及词性标注。THULAC 是一套中文词法分析工具包，集成了世界上最大规模的中文语料库，中文分词及词性标注能力强、准确率高、速度快[20]。

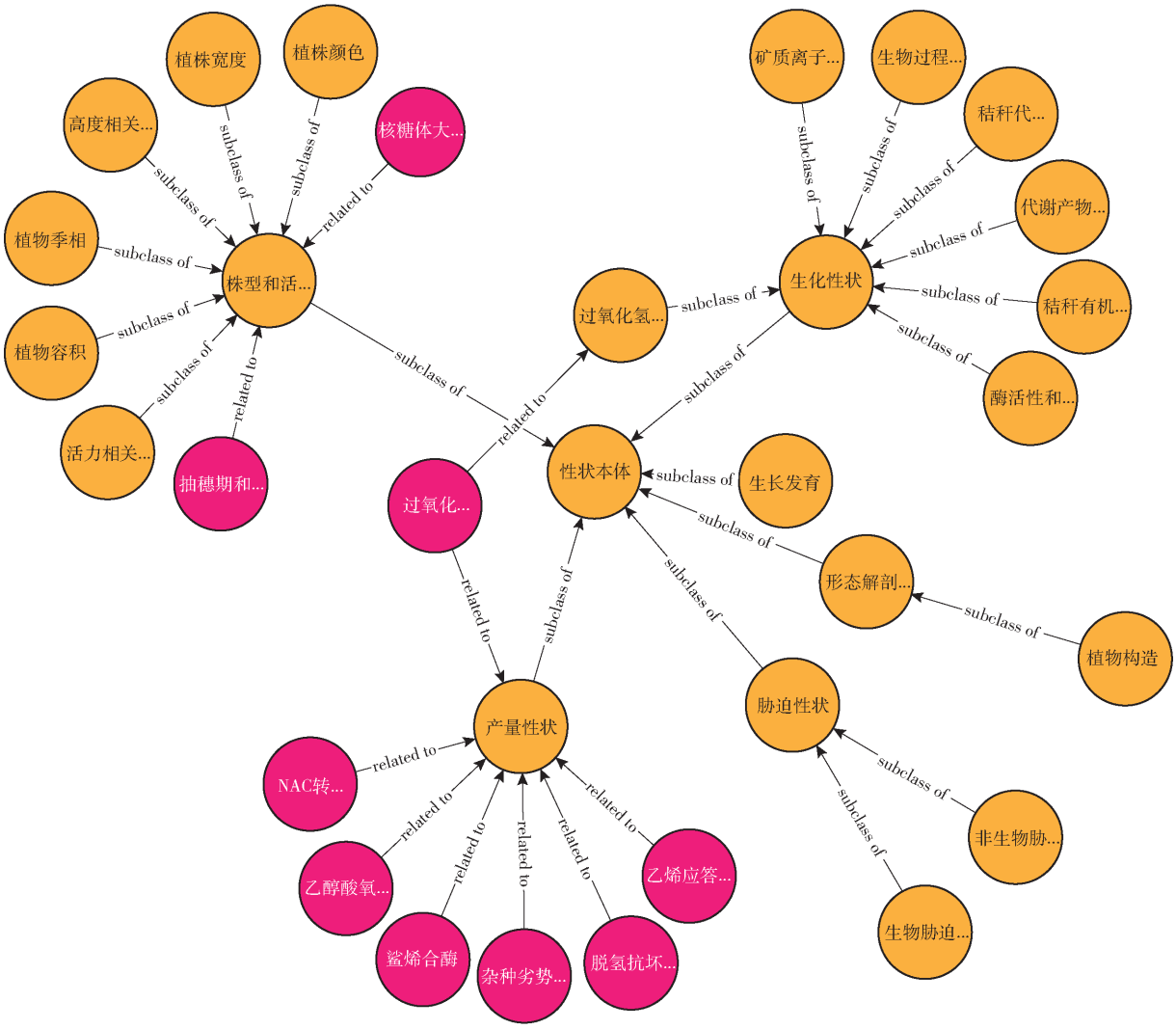

图 1 水稻表型组学实体-关系实例

Fig. 1 Entity-relationship instances of rice phenomics

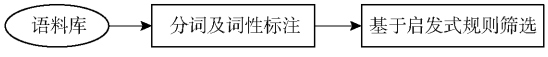

图 2 水稻表型组学实体识别过程

Fig. 2 Rice phenomics entities recognition process

启发式规则是指在解决问题时根据以往经验而制定的一系列规则,重点在于根据经验选择行之有效的方法[21]。本文水稻表型组学实体启发式规则如图 3 所示。

水稻表型组学实体识别中,根据已有经验,表型组学实体词性只能为名词,因此除了名词之外的其他词性应被丢弃。在名词基础之上应该检验该词的前一词是否为修饰性词性,是则与该词组成名词性短语。

图 4 为本文所列实体识别结果的示例,可以看出“小肠”、“叶序”等词属于满足当前词性要求的词语,“镁离子浓度”、“植物细胞长度”等词属于当前词与其前一词的组合词语。由于水稻表型术语与基因组术语是专业性实体名词,因此本文对水稻表型术语及基因组术语不再进行分词及词性标注,直接

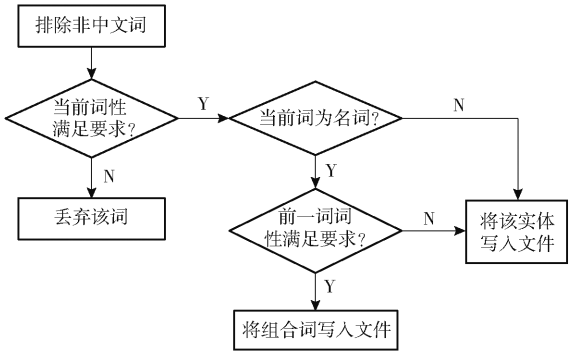

图 3 水稻表型组学实体判断的启发式规则

Fig. 3 Heuristic rules for determination of rice phenomics entities

将这两类词语加入到实体识别结果中,最终共 6 445 个词。

2 水稻表型组学实体分类

2.1 水稻表型组学数据预处理

对水稻表型组学语料库所做的预处理操作为将文本处理成特征矩阵。文本向量化是将文本特征表

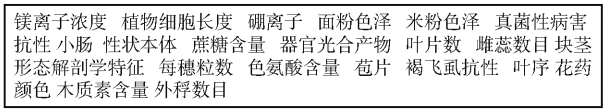

图 4　水稻表型组学实体示例

Fig. 4　Example of rice phenomics entities

示成一系列计算机可以计算和理解的，能够表达文本语义信息的向量形式[22]，通过一定的数理统计方法和计算机技术展现文本特征。目前文本向量化大部分是通过词向量化实现。

本文使用 TF－IDF 技术并结合 LSI 模型对数据进行预处理，计算文本词频特征，解决文本语义问题，对水稻文本数据进行特征向量化形成语义矩阵。

如图 5 所示，该过程首先将文本语料字典转换成词袋向量，便于计算其 TF－IDF 值，将转换成的 TF－IDF 向量转换成 LSI 向量，接着将生成的向量转换成稀疏矩阵，稀疏向量最后转换成密集向量，即为机器学习分类器模型接受的输入形式。

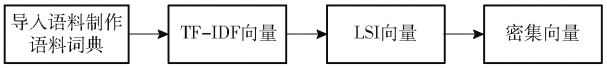

图 5　水稻表型组学数据预处理过程

Fig. 5　Preprocessing of rice phenomics data

2.1.1　TF－IDF 模型

词袋(Bag of word, BOW)[23]模型是最早出现的文本向量化方法，它将词语作为文本的基本组成单元，该模型与每个词语在文本中出现的频率具有一定的相关性，其产生的向量与文本词语出现的顺序无关。该模型原理比较简单，但易出现维度灾难和语义鸿沟等问题。

TF－IDF(Term frequency－inverse document frequency)[24]是一种基于词袋模型改进产生的，以计算词频及逆文本频率来计算词语分类能力，用于文本特征分析挖掘的常用方法。采用字词统计方法，分别统计一个文本中该词出现的频率和整个语料字典中该词出现的频率来综合评估字词的分类能力。TF 指词频，即一个字词在文本中出现的频率，出现次数越多，频率越高，TF 值就越高。IDF 指逆文本频率，若一个字词在某一类文本中出现的频率高，而在其他类别文本中出现的频率低，说明该字词具有良好的类别鉴别能力，对应该字词的 IDF 值越高。TF、IDF 的计算公式为

$$T_F(w)=\frac{n(w)}{S} \tag{1}$$

式中　$T_F(w)$——词 w 在当前文本中的词频

$n(w)$——词 w 在文章中出现的个数

S——文本的总词数

$$I_{DF}(w)=\lg\frac{N}{N(w)} \tag{2}$$

式中　$I_{DF}(w)$——词 w 在当前文本的逆文本频率

N——语料库中的文本总数

$N(w)$——语料库中包含词 w 的文本总数

TF－IDF 值的计算公式为

$$T_{F-IDF}(w)=T_F(w)I_{DF}(w) \tag{3}$$

以“半矮秆基因”URL 网页(http://www.ricedata.cn/gene/list/30.htm)为例，该文有 902 个词，“水稻”、“半矮秆基因”分别出现 13、2 次，通过式(1)得出词频(TF)为 0.014 4、0.002 2。网页总数为 1 966，包含“水稻”的网页数为 1 731 个，包含“半矮秆基因”的网页数为 16 个。则它们的逆文本频率 IDF 和 TF－IDF 计算示例如表 1 所示。

表 1　IDF 和 TF－IDF 计算示例

Tab. 1　Examples of IDF and TF－IDF evaluation

词语	含该词文本数	IDF	TF－IDF
水稻	1 731	lg(1 966/1 731) = 0.055 3	0.055 3 × (13/902) = 0.000 8
半矮秆基因	16	lg(1 966/16) = 2.089 5	2.089 5 × (2/902) = 0.004 6

表 1 中的 TF－IDF 值表明，常出现的词语“水稻”，应该给予较低权重。对于词语“半矮秆基因”，几乎没有在其他文本中出现，在出现的文本中它对应的 IDF 值较高，具有较强的类别鉴别能力。

2.1.2　LSI 模型

潜在语义模型(Latent semantic index, LSI)[25]为解决文本语义问题，如果文本中的某个词与其他词的相关性不大，则很可能是一词多义的问题。LSI 模型使用奇异值分解方法来分解词语-文本矩阵，将原始矩阵映射到语义空间内，形成语义矩阵。

利用奇异值分解方法降低矩阵维度，计算公式为

$$\boldsymbol{A}_{r\times c}=\boldsymbol{U}_{r\times j}\boldsymbol{\Sigma}_{j\times j}\boldsymbol{V}^{\mathrm{T}}_{j\times c} \tag{4}$$

式中　$\boldsymbol{A}_{r\times c}$——第 r 个文本第 c 个词的特征值

$\boldsymbol{U}_{r\times j}$——第 r 个文本第 j 个主题的相关度

$\boldsymbol{\Sigma}_{j\times j}$——第 j 个主题第 j 个词义相关度

$\boldsymbol{V}^{\mathrm{T}}_{j\times c}$——第 j 个词第 c 个词义的相关度

这里使用的特征值是基于预处理后的标准化 TF－IDF 值。j 是设定的特征维度，也称为主题数，一般比文本要少，一般设定在 100～200 之间，本文 j 设定维度为 150 维。经过一次奇异值分解，可以得到 r 个文本与 c 个词的关联程度。

2.2　水稻表型组学实体分类算法

2.2.1　Stacking 集成学习模型

基于 Stacking[26]集成学习策略是一种异构分类器集合的技术，被认为是实现集合中基分类器多样

性的工具,以此提高组合分类的准确性。它采用两层框架的结构,如图 6 所示。具体的训练过程为:首先是 Stacking 集成学习方法调用不同类型的分类器对数据集进行训练学习,然后将各分类器得到的训练结果组成一个新的训练样例作为元分类器的输入,最终第 2 层模型中元分类器的输出结果为最终的结果输出[27]。

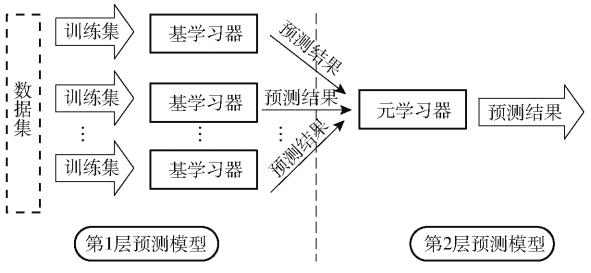

图 6 基于 Stacking 的集成学习方式

Fig. 6 Ensemble learning method based on Stacking

在组合模型中,对基分类器的选取有两种选择:一种是选取同种类型的基分类器,另一种是选取不同类型的基分类器,或者说是异质的。本文采用第 2 种选择方法,基于 Stacking 集成学习的两层结构框架对几种不同的分类器组合学习。具体使用 SVM、K-NN、GBDT 和 RF 作为第 1 层的分类器,即基分类器,RF 作为第 2 层的元分类器。针对这 4 种常见的单一分类器,具有性能优异且训练机理差距大的特点,提出了将这 4 种算法进行组合学习,从而进一步提高分类器的分类性能。

2.2.2 SVM 算法

SVM[28-29]是指一系列以支持向量机算法为基础的传统机器学习方法,可以用于文本的回归和分类问题。SVM 的主要原理是找到一个能够将所有数据样本划分开的超平面,使得样本集中的所有数据到这个超平面的距离最短。

使用 SVM 做分类问题主要分为线性分类和非线性分类。当一个分类问题是线性可分时,可以直接用一条直线将不同类别的数据进行区分,这条分类直线称为决策面。为求得最优的分类效果和性能,只需要计算数据与决策面的最大距离即可,寻找这个最大间隔的过程就叫最优化。但是,随着数据量的增多,数据分布不均衡,数据变成线性不可分的,这时就需要核函数将数据映射到特征空间,利用超平面对数据进行分类分割。

2.2.3 K-NN 算法

K-NN 算法[30]是一种通过计算不同数据之间特征值的距离进行分类的方法,是文本分类和回归常用的分类算法之一,目前对 K-NN 算法的理论研究已经非常充足。K-NN 算法的主要原理是,如果一个预测样本在特征空间内存在 K 个最近邻,那么预测样本的类别通常由 K 个近邻的多数类别决定。K-NN 算法中通过选择 K 值、距离度量方法,以及分类决策规则来优化算法的效果和性能。

K 值的选择一般根据样本的分布,通过交叉验证的方式选择一个最优的 K 值。若选择较小的 K 值,训练误差会减小,但容易使得训练的泛化误差增大,使模型复杂且容易过拟合。若选择较大的 K 值,可以减小泛化误差,模型预测结果一般不会出现过拟合现象,但会使训练误差增大,造成预测发生错误。因此,需要选择合适的 K 值训练模型,本文算法参数优化选择使用 GridSearch 进行验证。

2.2.4 GBDT 算法

GBDT 算法[31-32]提升分类器性能的方式不是为错分数据分配更高权重,而是直接修改残差计算的方式,原始的 Adaboost[33]限定了损失函数为平方损失函数,所以可直接进行残差计算。GBDT 通过定义更多种类的损失函数,采用负梯度的方式近似求解,这也是称其为梯度提升树的原因。由于这个改进,GBDT 不仅可以用于分类,也可以用于回归,有着更好的适应性。并且 GBDT 的非线性变换较多,所以其表达能力很强。

2.2.5 RF 算法

RF 算法[34]是一种集成算法,通过训练多个弱分类器集成一个强分类器,预测结果通过多个弱分类器的投票或取平均值决定,使得整体模型的结果具有较高的精确度和泛化性能。随机森林算法的主要原理是,通过随机采样方式构建一个随机森林,随机森林的组成单元是决策树,当进行预测时,森林中的每一棵决策树均参与分类预测,最终通过决策树投票选取得票数最高的分类作为预测分类。

2.3 水稻表型组学实体分类组合模型

考虑基学习器的预测能力,本文在 Stacking 集成学习的分类器组合模型第 1 层选择学习能力较强和差异度较大的模型作为基学习器,有助于模型整体的预测效果提升。其中,随机森林和梯度提升决策树分别采用 Bagging 和 Boosting 的集成学习方式,具有出色的学习能力和严谨的数学理论,在各个领域得到广泛应用。支持向量机对于解决小样本和非线性以及高纬度的回归问题具有优势。K-NN 算法因其理论成熟和训练高效等特点,有着良好的实际应用效果。第 2 层元学习器选择泛化能力较强的模型,用于纠正多个学习算法对于训练集的偏置情况,并通过集合方式防止过拟合效应出现。综上所述,Stacking 集成学习的分类器组合模型选择第 1 层基学习器分别为 SVM、K-NN、GBDT、RF,第 2 层元学

习器选择 RF。Stacking 集成学习的分类器组合模型如图 7 所示。

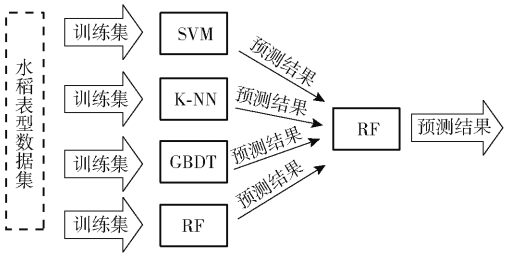

图 7　基于 Stacking 集成学习的水稻表型组学实体分类器组合模型

Fig. 7　Combination model of Stacking ensemble learning classification algorithm for rice phenomics entities

在水稻表型组学实体分类组合模型框架中，过程分为 5 步，处理过程总结为：

输入：训练集 $D=\{(y_i,\boldsymbol{x}_i);i=1,2,\cdots,m\}$，基分类器，测试集 T。

输出：测试集上的分类结果。

(1)使用 TF－IDF 技术结合 LSI 模型进行数据预处理。

(2)第 1 层预测算法的 K 个基学习器 M_k。

for $k=1$ to K

基于数据集 D_{-k} 训练第 1 层的基学习器 M_k

end

(3)构成新数据集

$D_{\text{new}}=\{(y_i,\boldsymbol{z}_{1i},\boldsymbol{z}_{2i},\cdots,\boldsymbol{z}_{ki});i=1,2,\cdots,m\}$

(4)第 2 层使用随机森林算法 M_{new} 预测模型训练。

基于 D_{new} 训练模型 M_{new}。

(5)使用测试集 T 分类，评价分类结果。

首先利用 TF－IDF 技术结合 LSI 模型的方法对获取的水稻表型组学数据进行文本预处理，预处理后的水稻表型组学数据 $D=\{(y_i,\boldsymbol{x}_i);i=1,2,\cdots,m\}$，$\boldsymbol{x}_i$ 为第 i 个样本的特征向量，y_i 为第 i 个样本对应的预测值，将数据集切分为训练集和测试集，随机将数据划分成 K 个大小基本相等的子集 $D_1,D_2,\cdots,D_k$。$D_{-k}=D-D_k$，定义 D_k 和 D_{-k} 分别为 K 折交叉验证中的第 k 折测试集与训练集。第 1 层包含 K 个基学习器，基学习器包括 SVM、GBDT、RF、K-NN。对训练集 D_{-k} 用第 k 个算法训练得到基模型 M_k，$k=1,2,\cdots,K$。对于 K 折交叉验证中的第 k 折测试集 D_k 中的每一个样本 $\boldsymbol{x}_i$，基学习器 M_k 对它的预测表示为 $\boldsymbol{z}_{ki}$。

在完成交叉验证过程后，将每个基学习器的各自预测结果构成新的数据样本，即：$D_{\text{new}}=\{(y_i,\boldsymbol{z}_{1i},\boldsymbol{z}_{2i},\cdots,\boldsymbol{z}_{ki}),i=1,2,\cdots,m\}$，重构特征，将第 1 层基学习器的输出结果作为第 2 层学习器 M_{new} 的输入，本文选择随机森林算法作为第 2 层预测模型的元学习器进行训练，再由第 2 层的元学习器模型输出最终预测结果，基于 Stacking 集成学习的分类器组合模型通过多个学习器的输出结果提升模型的分类能力，以获得整体预测精度的提升。

3　分类结果与分析

3.1　实验环境

本文实验环境：Intel(R) Core(TM) i5－8250U 1.6 GHz 处理器，8 GB 内存，500 GB 硬盘，64 位 Windows 10 操作系统。

3.2　参数设置

Stacking 集成学习通过将 K-NN、RF、SVM、GBDT 模型作为初级学习器，4 个验证集的输出组合成次学习器的一个输入特征，进行再次训练。各算法的参数设置如下：SVC 算法模型使用 RBF 核函数，gamma 参数设置为 0.1，惩罚系数 C 设置为 3；K-NN 算法模型近邻个数设置为 6，用 L_2 距离来标识每个样本的近邻样本权重，权重与距离成反比；RF 算法模型计算属性的信息增益值来选择合适的结点，选择划分的属性值最大值为 9，叶子结点的最少样本数为 2，决策树的个数为 81；GBDT 算法模型的弱学习数量为 100，学习速率设置为 0.1，回归树的深度为 3。训练集和测试集的大小比例设置为 7∶3。

3.3　数据集

基于 Stacking 集成学习的分类器组合模型，要实现水稻表型组学实体自动分类，需利用人工分类方法制作数据集，数据集的品质直接影响到机器训练的品质。在进行人工分类时，首先对含有明显特定字词的词语进行提取，对不明显的词语通过人工分类的方式逐词进行标注[35]。水稻表型组学实体分类情况参照文献[36]，如表 2 所示。

表 2　水稻表型组学实体分类情况

Tab. 2　Detail of entity classification of rice phenomics

类别	类别说明	实体数量
Gene	基因组，包括基因、蛋白质、酶等	1 389
Chemical	化学环境，包括营养素、维生素、植物激素、肥料、农药、杀菌剂、矿物质等	262
Environment	自然环境，包括阳光、水、真菌、细菌、媒介等	213
Pathology	水稻生理	77
Phenotype	水稻表型性状，包括表型术语、具体表现等	1 221
Other	其他实体，包括人名、地名、机构名、植物名词等	1 595

3.4　算法性能评估指标

本文使用的 5 种机器学习方法：基于 Stacking

集成学习的分类器组合模型、SVC 算法、K-NN 算法、RF 算法、GBDT 算法均使用 TF - LSI 方法预处理，使用 GridSearch 进行算法模型参数调优。实验模型的评估使用精确率 P(Precision)、召回率 R(Recall)、综合评价指标 F_1(F1 - Measure)、准确率 A(Accuracy)等指标，计算公式为

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{5}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{6}$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{7}$$

$$A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \tag{8}$$

式中 T_P——将实际为正类预测为正类的数量

T_N——将实际为正类预测为负类的数量

F_P——将实际为负类预测为正类的数量

F_N——将实际为负类预测为负类的数量

F_1 是模型精确率和召回率的一种加权平均值，介于 0 到 1 之间，F_1 越高，代表分类器的综合性能越好。F_1 是分类问题中的一个常用的衡量指标，能够对模型精确率与召回率进行综合评估。

3.5 Stacking 算法分类结果

利用 TF - IDF 技术结合 LSI 模型的方法进行水稻表型组学数据预处理，作为 Stacking 集成学习的分类器组合模型的输入样例。这里以获取的水稻表型组学数据为例，首先对 Stacking 集成学习的分类器组合模型分类结果进行分析，然后对比单一分类算法，从而验证 Stacking 集成学习的分类器组合模型分类性能的优异性。

Stacking 集成学习吸收融合模型的优点提高准确率和稳定性。通过将 K-NN、RF、SVM、GBDT 模型作为初级学习器，4 个验证集的输出组合成次学习器的一个输入特征，进行再次训练。Stacking 集成学习的分类器组合模型实验结果如表 3 所示。

表 3 Stacking 算法各实体类别实验结果

Tab. 3 Experimental results of various kinds of rice entities with Stacking algorithm %

类别	精确度	召回率	F_1
Gene	90.94	90.00	90.47
Chemical	69.44	48.08	56.82
Environment	89.29	59.52	71.43
Pathology	46.67	46.67	46.67
Phenotype	76.47	78.45	77.45
Other	76.66	83.13	79.76

从表 3 结果来看，在水稻表型组学数据中，组合模型对数据量较大的分类如 Environment 类、Phenotype 类和 Other 类的分类精确度均达到 76% 以上，最佳的分类是 Gene 类，达到 90.9% 以上的分类精确度，多分类结果的召回率和 F_1 值均在 90% 以上。对于不平衡数据集，Stacking 集成学习的分类器组合模型在数据量较大的类别上的表现较为突出，在水稻表型组学数据的多分类问题上具有较强的分类性能。

3.6 分类结果对比

将对单一分类器算法和 Stacking 集成学习的分类器组合模型的分类结果进行对比分析。各分类算法对水稻表型组学实体的分类精度对比如图 8 所示。

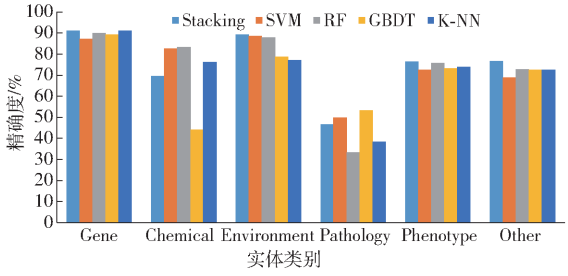

图 8 各分类算法精确度对比

Fig. 8 Comparison of precision of classification algorithms

由图 8 可以看出，Stacking 算法对各类别的分类效果均较好，包括类别数量较少的 Chemical 类和 Environment 类，整体上优于其他分类方法，对 Gene 类的分类精确度达 90.9%。

各分类算法对水稻表型组学实体的分类召回率对比如图 9 所示。

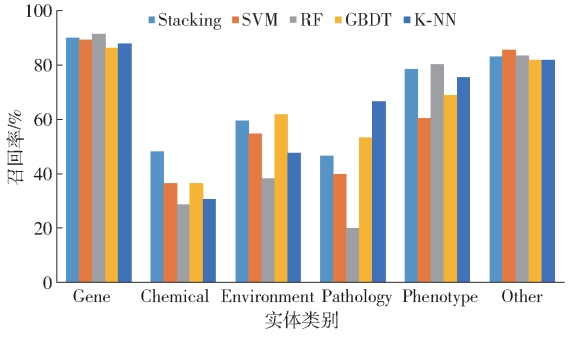

图 9 各分类算法召回率对比

Fig. 9 Comparison of recall rate of classification algorithms

由图 9 可以看出，水稻表型组学数据的 Gene 类、Phenotype 类、Other 类的整体召回率均较高，尤其是 Stacking 集成学习模型和随机森林算法在其 3 种分类上表现较优。但无论是针对哪种类型水稻表型组学实体，Stacking 集成学习模型分类结果的召回率均较高，显示其较强的分类性能。

各分类算法对水稻表型组学实体的分类 F_1 对比如图 10 所示。

由图 10 整体来看，Gene 类、Phenotype 类、Other

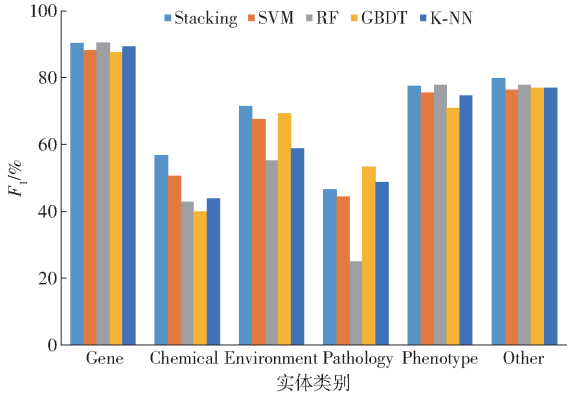

图 10　各分类算法 F_1 值对比

Fig. 10　Comparison of F1-Measure of classification algorithms

类 3 类算法整体分类的精确率和召回率的加权平均值 F_1 较高，Chemical 类、Environment 类和 Pathology 类的 F_1 相对较低，这是由于各类数据集数量分布不均衡造成的，相对 Gene 类、Phenotype 类和 Other 类训练数据较多，分类更精确，综合性能较好。

Stacking、RF、K-NN、GBDT、SVM 各分类算法准确率分别为 80.55%、72.97%、67.51%、76.24%、78.34%，通过对比 K-NN、GBDT、RF、SVM 4 种算法，Stacking 集成学习的分类器组合模型的准确率最高。Stacking 集成学习算法集成了其他 4 种弱分类器，对于准确率有一定的提升。相比单一分类器 K-NN 算法，基于 Stacking 集成学习的分类器组合模型的分类准确率高 13.04 个百分点，性能提升较大，平均比单一分类器提高 6.78 个百分点。

综上所述，在 TF－IDF 技术结合 LSI 模型方法进行水稻表型组学数据预处理的情况下，Stacking 集成学习的分类器组合模型在本文水稻表型组学数据集上的整体效果最佳。对于不平衡数据集，Stacking 集成学习分类器的组合模型的表现较为突出，整体精确度、召回率和 F_1 值均较高。总体上，Stacking 集成学习的分类器组合模型较单一的分类器，在分类性能上有一定的提升，准确率较分类性能较好的 SVM 算法提高 2.21 个百分点，较 K-NN 算法准确率提高 13.04 个百分点。

4　结束语

针对水稻表型组学数据，采用 TF－IDF 技术和潜在语义模型进行预处理，基于堆叠式两阶段集成学习的分类器组合模型，结合 K-近邻算法、支持向量机、随机森林，梯度提升决策树机器学习方法，提升水稻表型组学实体数据分类的性能，对于不平衡的水稻表型组学数据集和对不同类别的水稻表型组学数据都具有较好的多分类能力，总体准确率为 80.55%，相对 SVC、K-NN、GBDT、RF 单一分类器，表现最佳，分类准确率平均高 6.78 个百分点。

参　考　文　献

[1] 武开阔，张丽莉，宋玉超，等. 稳定性氮肥配合秸秆还田对水稻产量及 N_2O 和 CH_4 排放的影响[J]. 应用生态学报，2019，30(4)：1287－1294.
WU Kaikuo，ZHANG Lili，SONG Yuchao，et al. Effects of stabilized N fertilizer combined with straw returning on rice yield and emission of N_2O and CH_4 in a paddy field[J]. Chinese Journal of Applied Ecology，2019，30(4)：1287－1294.（in Chinese）

[2] 段凌凤，杨万能. 水稻表型组学研究概况和展望[J]. 生命科学，2016，28(10)：1129－1137.
DUAN Lingfeng，YANG Wanneng. Research advances and future scenarios of rice phenomics[J]. Chinese Bulletin of Life Sciences，2016，28(10)：1129－1137.（in Chinese）

[3] 唐惠燕，倪峰，李小涛，等. 基于 Scopus 的植物表型组学研究进展分析[J]. 南京农业大学学报，2018，41(6)：1133－1141.
TANG Huiyan，NI Feng，LI Xiaotao，et al. Analysis of the advance in plant phenomics research based on Scopus tools [J]. Journal of Nanjing Agricultural University，2018，41(6)：1133－1141.（in Chinese）

[4] PIERUSCHKA R，POORTER H. Phenotyping plants：genes，phenes and machines introduction[J]. Funct Plant Biol，2012，39(11)：813－820.

[5] TESTER M，LANGRIDGE P. Breeding technologies to increase crop production in a changing world[J]. Science，2010，327(5967)：818－822.

[6] WHITE J W，ANDRADE-SANCHEZ P，GORE M A，et al. Field-based phenomics for plant genetics research[J]. Field Crops Res.，2012，133：101－112.

[7] YANG W，DUAN L，CHEN G，et al. Plant phenomics and high-throughput phenotyping：accelerating rice functional genomics using multidisciplinary technologies[J]. Curr. Opin. Plant Biol.，2013，16(2)：180－187.

[8] 李涓子，侯磊. 知识图谱研究综述 [J]. 山西大学学报(自然科学版)，2017，40(3)：454－459.
LI Juanzi，HOU Lei. Reviews on knowledge graph research[J]. Journal of Shanxi University(Natural Science Edition)，2017，40(3)：454－459.（in Chinese）

[9] ALTHUWAYNEE O F，PRADHAN B，PARK H J，et al. A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping [J]. Catena，2014，114：21－36.

[10] MOHAMMADI E. Knowledge mapping of the Iranian nanoscience and technology：a text mining approach[J]. Scientometrics，

2012, 92(3): 593 - 608.
[11] YAO Q, CHEN K, YAO L, et al. Scientometric trends and knowledge maps of global health systems research[J]. Health Research Policy and Systems, 2014, 12(26): 1 - 20.
[12] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582 - 600.
LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582 - 600. (in Chinese)
[13] KRAWCZYK B, MINKU L L, GAMA J, et al. Ensemble learning for data stream analysis: a survey[J]. Information Fusion, 2017, 37: 132 - 156.
[14] RAYHAN F, AHMED S, MAHBUB A, et al. MEBoost: mixing estimators with boosting for imbalanced data classification[C]// 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). IEEE, 2017: 1 - 6.
[15] WANG B, PINEAU J. Online bagging and boosting for imbalanced data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3353 - 3366.
[16] XIA Y, LIU C, LI Y Y, et al. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring [J]. Expert Systems with Applications, 2017, 78: 225 - 241.
[17] CZARNOWSKI I, JEDRZEJOWICZ P. An approach to machine classification based on stacked generalization and instance selection[C]// 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2016. 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016: 004836 - 004841.
[18] 李乔宇, 尚明华, 王富军, 等. 基于 Scrapy 的农业网络数据爬取[J]. 山东农业科学, 2018, 50(1): 142 - 147.
LI Qiaoyu, SHANG Minghua, WANG Fujun, et al. Data crawling from agricultural internet based on Scrapy[J]. Shandong Agricultural Sciences, 2018, 50(1): 142 - 147. (in Chinese)
[19] WEBBER J, ROBINSON I. A programmatic introduction to Neo4j[M]. Addison-Wesley Professional, 2018.
[20] LI Z, SUN M. Punctuation as implicit annotations for Chinese word segmentation[J]. Computational Linguistics, 2009, 35(4): 505 - 512.
[21] 卢建云, 朱庆生, 吴全旺. 一种启发式确定聚类数方法[J]. 小型微型计算机系统, 2018, 39(7): 7 - 11.
LU Jianyun, ZHU Qingsheng, WU Quanwang. Heuristic method of determining the number of clusters[J]. Journal of Chinese Computer Systems, 2018, 39(7): 7 - 11. (in Chinese)
[22] CHA J, KIM J, PARK S. GRiD: gathering rich data from PubMed using one-class SVM[C]// IEEE International Conference on Systems. IEEE, 2017.
[23] ZHAO R, MAO K. Fuzzy bag-of-words model for document representation[J]. IEEE Transactions on Fuzzy Systems, 2017, 26(2): 794 - 804.
[24] DAS B, CHAKRABORTY S. An improved text sentiment classification model using TF - IDF and next word negation[J]. arXiv preprint arXiv:1806.06407, 2018.
[25] 崔彤彤, 崔荣一. 基于潜在语义分析的文本指纹提取方法[J]. 中文信息学报, 2018, 32(5): 79 - 84.
CUI Tongtong, CUI Rongyi. Text fingerprint extraction based on latent semantic analysis[J]. Journal of Chinese Information Processing, 2018, 32(5): 79 - 84. (in Chinese)
[26] WOLPERT D H. Stacked generalization[M]. Los Alamos: Stacked Generalization, 2017:241 - 259.
[27] MEHMOOD A, ON B W, LEE I, et al. Spam comments prediction using stacking with ensemble learning[C]// Journal of Physics: Conference Series. IOP Publishing, 2018, 933(1): 012012.
[28] DIREITO B, TEIXEIRA C A, SALES F, et al. A realistic seizure prediction study based on multiclass SVM[J]. International Journal of Neural Systems, 2017, 27(3): 470 - 472.
[29] GURURAJAPATHY S S, MOKHLIS H, LLLIAS H A B, et al. Support vector classification and regression for fault location in distribution system using voltage sag profile[J]. IEEJ Transactions on Electrical and Electronic Engineering, 2017, 12(4): 519 - 526.
[30] ZHANG S, LI X, ZONG M, et al. Learning k for K-NN classification[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2017, 8(3): 43.
[31] DE MELO V V, USHIZIMA D M, BARACHL S F, et al. Gradient boosting decision trees for echocardiogram images[C]// 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1 - 8.
[32] MA X, DING C, LUAN S, et al. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(9): 2303 - 2310.
[33] MINZ A, MAHOBIYA C. MR Image classification using adaboost for brain tumor type[C]// 2017 IEEE 7th International Advance Computing Conference (IACC). IEEE, 2017: 701 - 705.
[34] THANH NOI P, KAPPAS M. Comparison of random forest, K-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel - 2 imagery[J]. Sensors, 2018, 18(1): 1 - 20.
[35] 王蕾, 谢云, 周俊生, 等. 基于神经网络的片段级中文命名实体识别[J]. 中文信息学报, 2018, 32(3): 84 - 90,100.
WANG Lei, XIE Yun, ZHOU Junsheng, et al. Segment-level Chinese named entity recognition based on neural network[J]. Journal of Chinese Information Processing, 2018, 32(3): 84 - 90,100. (in Chinese)
[36] 潘映红. 论植物表型组和植物表型组学的概念与范畴[J]. 作物学报, 2015, 41(2): 175 - 186.
PAN Yinghong. Analysis of concepts and categories of plant phenome and phenomics[J]. Acta Agronomica Sinica, 2015, 41(2): 175 - 186. (in Chinese)