doi:10.6041/j.issn.1000-1298.2021.09.026

基于 Shapley 值组合预测的玉米单产估测

王鹏新[1]　乔　琛[1]　李　俐[2]　周西嘉[1]　许连香[2]　胡亚京[1]

(1. 中国农业大学信息与电气工程学院, 北京 100083; 2. 中国农业大学土地科学与技术学院, 北京 100083)

摘要: 为进一步促进机器学习技术在玉米单产估测中的应用,以河北中部平原为研究区域,选取与玉米长势和产量密切相关的条件植被温度指数(Vegetation temperature condition index,VTCI)和叶面积指数(Leaf area index,LAI)为特征变量,通过极限梯度提升(Extreme gradient boosting,XGBoost)算法和随机森林(Random forest,RF)算法分别对玉米单产进行估测。基于组合预测思想与 Shapley 值理论,分别确定组合预测模型中 XGBoost 与 RF 模型权重,进而得到组合预测模型,结果表明,基于 Shapley 值确定的组合估产模型精度较高(R^2 = 0.32),达极显著水平(P < 0.001)。同时将组合预测模型应用于河北中部平原 2012 年各县(区)玉米的单产估测,结果表明,模型精度较高(R^2 = 0.52),玉米估测单产与实际单产的平均相对误差和均方根误差分别为 9.86%、831.14 kg/km^2,达到极显著水平(P < 0.001),且组合预测模型的精度均优于单一估测模型。研究发现,河北中部平原玉米估测单产随年份发生波动变化,呈先降低后升高的趋势。玉米估测单产以西部地区最高,其次是北部和南部地区,东部地区最低。

关键词: 玉米; 估产; 条件植被温度指数; 叶面积指数; Shapley 值; 组合预测

中图分类号: TP79　　**文献标识码**: A　　**文章编号**: 1000-1298(2021)09-0221-09　　**OSID**:

Estimation of Maize Yield Based on Shapley Value Combination Forecasting

WANG Pengxin[1]　QIAO Chen[1]　LI Li[2]　ZHOU Xijia[1]　XU Lianxiang[2]　HU Yajing[1]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *College of Land Science and Technology*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: Aiming to promote the application of machine learning in agriculture field and improve accuracy of the maize yield estimation, the central plain of Hebei Province was selected as the study area, which includes fifty-three counties (districts). Vegetation temperature condition index (VTCI) and leaf area index (LAI) at the main growth stages of maize were selected as key crop growth indicators for estimating the maize yield by using two machine learning methods, extreme gradient boosting (XGBoost) and random forest (RF), and as well as their combination. Firstly, the XGBoost and RF were used to estimate yield of maize from 2010 to 2017, then the XGBoost and RF's weights were determined by combination forecasting model by using the Shapley value method, and finally maize yield of each county in 2012 was estimated based on the combination forecasting model. The results showed that the mean relative error (MRE) and root mean square error (RMSE) between the estimated yield of maize and the actual yield were 9.86% and 831.14 kg/km^2, respectively. The accuracy of the combination forecasting model (R^2 = 0.52, P < 0.001) was better than that of the XGBoost model and RF model, which can be applied to estimate the yield of maize in the study area. The combination model was used to estimate the maize yield of the central plain of Hebei Province pixel by pixel from 2010 to 2018. The estimated yield of maize showed a trend of decrease first and then increase over time. The spatial distribution of maize yield was the highest in the western region, followed by the northern and southern regions, and the eastern region was the lowest. The results showed that the temporal and spatial changes of maize in the central plain of Hebei Province were in line with reality, and the research result can provide guidance for the growth monitoring and yield estimation of maize in the study area.

Key words: maize; estimation of yield; vegetation temperature condition index; leaf area index; Shapley value; combination forecasting

收稿日期: 2020-09-18　修回日期: 2020-10-15

基金项目: 国家重点研发计划项目(2016YFD0300603-3)

作者简介: 王鹏新(1965—),男,教授,博士,主要从事定量遥感及其在农业中的应用研究,E-mail: wangpx@cau.edu.cn

通信作者: 李俐(1976—),女,副教授,博士,主要从事微波遥感及其在农业中的应用研究,E-mail: lilixch@cau.edu.cn

0 引言

粮食的生产与安全对提升我国农业的经营管理水平、完善农作物的种植结构和确保粮食安全等具有重要意义[1-3],因此及时准确估测农作物的产量有利于维持国家稳定和促进经济发展。

干旱作为影响农作物生长发育及产量的重要因素已受到广泛的关注[4-5]。王鹏新等[6]在归一化植被指数(Normalized difference vegetation index, NDVI)和地表温度(Land surface temperature, LST)的散点图呈三角形区域分布的基础上,提出了条件植被温度指数(Vegetation temperature condition index, VTCI)的干旱监测方法,可用于反映农作物生长过程中的水分亏缺信息。孙威等[7]对 VTCI 冷热边界的确定方法进行了完善,并验证了利用 VTCI 进行干旱监测的可行性。农作物产量不仅受到水分胁迫的影响,还与作物的生长状态密切相关[8],叶面积指数(Leaf area index, LAI)通过对作物的光合速率和干物质的积累量的反映能较好地表征农作物的生长状态[9]。

遥感技术对比传统统计调查方法,凭借其覆盖范围广、重访周期短等独特优势被广泛应用于作物长势监测及产量估测,同时对大规模的农业生产调查、评价、监测和管理具有独特的作用[10]。随着大数据技术的发展,遥感大数据将推动农业遥感估产的发展[11]。目前,遥感估产方法中常采用的统计模型、机理模型和半机理模型等均能够较好地对作物进行估产[12]。但由于在实际应用中机理和半机理模型存在需要输入较多的参数问题,因此机理和半机理模型存在一定的局限性。而统计模型其估产的精确度依赖于选取遥感影像的时相,对作物的生长和产量形成的机理解释性不强[12],因此在实际应用中同样具有一定局限性。在作物生长过程中,经常受到各种因素影响,同时,这些因素在作物不同生育时期产生不同的影响,即使采用相同的估算方式也会得到不同的结果与估测精度。虽然其估产精度不同,但是这些模型可以提供不同的有用信息,如将一些方法丢弃,就会失去有用信息,从而影响估产的精度[13]。因此,可以将单个估测模型提供的有用信息进行组合,以提高估产的精度。组合预测(Combination forecasting, CF)将不同预测模型进行有效组合,可视为对无限逼近真实数据生成过程的有效补充[14]。

BATES 等[15]首次提出了组合预测理论,这一预测理论在提高精度的同时更充分利用了预测样本所表达的信息,受到了国内外广大学者重视。组合预测根据各单一模型的信息贡献度,进而基于计算得到的单一模型权重构建组合预测模型,从而实现减少预测误差、提高预测精度的目标[16]。特别是在时间序列中分析真实数据生成过程中,通常具有区制转换(Regime shift)或参数漂移等特性。组合预测方法的引入,可减少由参数错误或模型错误带来的预测误差[17],甚至在单一预测结果存在有偏性的情况下,通过组合能产生具有无偏性的预测结果[18]。因此,组合模型具有普遍性的优点,最终的预测结果更接近实际数据。根据组合预测方法综合手段的不同,可分为权重综合法和区域综合法两种。与区域综合法相比,权重综合法得到了广泛的应用。目前,组合预测现在已经在多个领域内得到了应用[19-20],但是在农业领域还少有报道。

本文以河北中部平原为研究区域,选取与玉米长势和产量密切相关的 VTCI 和 LAI 为特征变量,采用极限梯度提升树(Extreme gradient boosting, XGBoost)和随机森林(Random forest, RF)两种机器学习算法模型分别估测研究区域的玉米单产,并借鉴经济学中合作博弈论 Shapley 值利益分配理论,通过组合预测模型中的权重合成思想,以单一预测模型的均方误差为基础确定单一模型权重得到组合预测模型,以期为玉米长势监测和产量估测提供新方法。

1 材料与方法

1.1 研究区概况

河北中部平原属黄淮海平原组成部分,是我国重要的玉米种植和生产基地[21],位于 114°32′~117°36′E 与 36°57′~39°50′N 之间,包括石家庄等 5 个市的 53 个县(区),其土地面积约为 5.30×10^4 km^2。河北中部平原地处典型温带大陆性季风气候区,四季分明、雨热同期。河北省中部平原年降水量 350~800 mm。降水时空分布不均,南方的降水量比北方高,夏季降水量高于冬季,丰水年降水量与枯水年降水量相差较大。该地区的耕种管理制度为一年两熟,结合该地区玉米实际生长状况,本文将河北中部平原玉米生长划分为 4 个生育时期:7 月上旬至 7 月中旬的出苗-拔节期、7 月下旬至 8 月上旬的拔节-抽穗期、8 月中旬至 9 月上旬的抽穗-乳熟期和 9 月中旬至 9 月下旬的乳熟-成熟期。根据王鹏新等[22]提出的作物分类方法,进而获得研究区域 2010—2018 年玉米种植区分布图,其中河北中部平原 2017 年玉米种植区分布图如图 1 所示。

1.2 数据获取与处理

1.2.1 VTCI 时间序列的生成

选取 2010—2018 年每年 7—9 月空间分辨率和

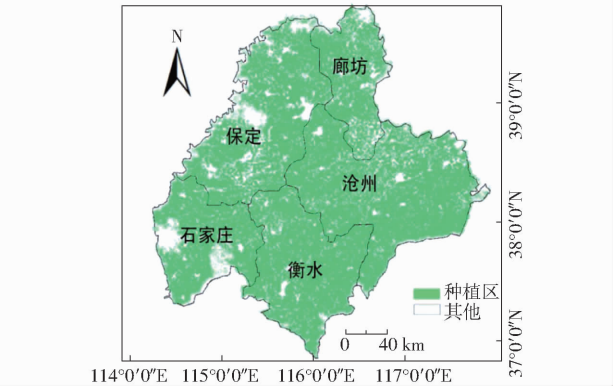

图 1　研究区域位置及玉米种植区(2017 年)

Fig. 1　Location of study area and planting area of maize(2017)

时间分辨率分别为 1 000 m、1 d 的 MODIS 日地表温度产品(MYD11A1)及日地表反射率产品(MYD09GA)。利用 MRT 对日地表温度和日地表反射率产品进行预处理之后,得到研究区域的日 LST 和日 NDVI 产品,运用最大值合成法生成旬 LST 与 NDVI 的最大值合成产品;基于生成的多年某一旬的 NDVI 和 LST 最大值合成产品再次使用最大值合成技术生成多年的旬 NDVI 和 LST 的最大值合成产品;基于生成的多年某一旬的 LST 最大值合成产品利用最小值合成技术逐像素提取最小值,计算得到多年旬 LST 的最大—最小值合成产品,并以此通过计算生成河北中部平原旬 VTCI 时间序列数据。VTCI 的计算公式为[6,23]

$$VTCI = \frac{LST_{\max}(NDVI_i) - LST(NDVI_i)}{LST_{\max}(NDVI_i) - LST_{\min}(NDVI_i)} \quad (1)$$

其中

$$NDVI = \frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + \rho_{red}} \quad (2)$$

$$LST_{\max}(NDVI_i) = a + bNDVI_i \quad (3)$$

$$LST_{\min}(NDVI_i) = a' + b'NDVI_i \quad (4)$$

式中　$NDVI$——归一化植被指数

ρ_{NIR}、ρ_{red}——近红外、红光波段的反射率

LST——地表温度

$LST_{\max}(NDVI_i)$、$LST_{\min}(NDVI_i)$——当 $NDVI_i$ 为某一特定值时所有像素地表温度最大值和最小值,即热边界和冷边界

$LST(NDVI_i)$——研究区域内某一像素的 $NDVI$ 值为 $NDVI_i$ 时的地表温度

a、b、a'、b'——LST 和 $NDVI$ 散点图近似得到的待定系数

1.2.2　LAI 时间序列的生成

选取 2010—2018 年 7—9 月的 MODIS 叶面积指数产品(MCD15A3H),其空间分辨率和时间分辨率分别为 500 m、4 d。利用上包络线 Savitzky－Golay 滤波对经 MRT 处理后得到的原始 LAI 产品进行平滑处理以消除云层覆盖、大气溶胶等因素引起的数据骤降的现象[23],滤波处理后的叶面积指数变化趋于平稳且更加符合玉米的生长物侯特征,解决了原始数据存在的质量问题。由于 LAI 与 VTCI 时间分辨率和取值范围不同,因此对滤波后的叶面积指数进行归一化处理,并通过观察多时相 MODIS 的 MCD15A3H 原始数据相元统计直方图将取值范围设置为 0～7,进而计算得到研究区域 2010—2018 年玉米主要生育时期的 LAI 时间序列数据。

1.2.3　玉米生育时期 VTCI 和 LAI 的计算及异常点数据处理

基于研究区玉米生育时期的划分结果,将生育时期内所包含的多旬 VTCI 和 LAI 平均值作为研究区域玉米该生育时期逐像素的 VTCI 和 LAI 值;通过叠加河北中部平原各县(区)行政边界图,将各县(区)所包含的玉米生育时期逐像素 VTCI 和 LAI 值的平均值作为生育时期县(区)尺度的 VTCI 和 LAI 值。同时在构建回归模型时,剔除加权 VTCI 与 LAI 残差置信区间在[－4 000,4 000] kg/km^2 以外的异常点数据。

1.3　研究方法

1.3.1　极限梯度提升算法

极限梯度提升算法(XGBoost)是一种基于梯度提升的决策树(Gradient boosting decision tree)集成算法[24],其基分类器主要包含分类和回归树(Classification and regression tree, CART)。本文玉米单产估测是回归问题,所以其基模型选择为回归树。含有 K 棵决策树 XGBoost 的树集成模型定义为

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (5)$$

式中　$\hat{y}_i$——样本 i 的预测值,即模型估测产量

x_i——样本所对应的特征变量 VTCI 与 LAI

f_k——第 k 个决策树的预测函数

树集成优化模型可以定义为

$$O_{bj} = \sum_{i=1}^{n} l(y_i, \hat{y}) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

其中

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| \boldsymbol{\omega}^2 \| \quad (7)$$

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

式中　$l(y_i, \hat{y}_i)$——损失函数,即均方误差

$\Omega(f)$——正则化项

γ——复杂度参数

T——树中叶子节点个数

λ——固定系数

$\boldsymbol{\omega}$——叶子节点量化权重向量

求解式(6)的优化问题,即求解 CART 树的结构。通过保留训练好的前 $t-1$ 轮树模型不变,在第 t 轮时添加一个新的预测函数,迭代计算得到最终预测结果[25]

$$\begin{cases}\hat{y}_i^{(0)}=0\\ \hat{y}_i^{(1)}=f_1(x_i)=\hat{y}_i^{(0)}+f_1(x_i)\\ \hat{y}_i^{(2)}=f_1(x_i)+f_2(x_i)=\hat{y}_i^{(1)}+f_2(x_i)\\ \quad\vdots\\ \hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i)\end{cases}\tag{8}$$

式中 $\hat{y}_i^{(t)}$——t 轮时目标量预测值

$\hat{y}_i^{(t-1)}$——保留的 $t-1$ 轮的目标量预测值

$f_t(x_i)$——第 t 轮加入的新的预测函数

XGBoost 模型可以通过输出玉米 4 个生育时期 VTCI 或 LAI 的特征变量重要性来评估不同特征变量对玉米产量的影响程度。在 XGBoost 中常用基于增益、覆盖度、频率的特征重要性指标进行特征重要性评价,其中基于增益(gain)表示各特征变量对 XGBoost 模型中每棵树采取每个特征变量的贡献而计算的模型相对贡献度为

$$G=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-c\tag{9}$$

式中 G——VTCI 与 LAI 基于增益获取的特征重要性得分值

G_L、G_R——左、右子树中的梯度

H_L、H_R——左、右子树中的二阶梯度

c——给定的临界值

1.3.2 随机森林算法

随机森林(RF)是一种基于袋装法(bagging)理论实现的集成学习算法,其基模型是决策树模型。随机森林具有极好的准确率,同时可以评估各特征在回归问题上的重要性。随机森林将若干没有联系的回归决策树$\{y(x,\theta_k),k=1,2,\cdots,K\}$构成 K 棵集成决策树[26],可以定义为

$$y(x)=\frac{1}{N}\sum_{k=1}^{K}h(x,\theta_k)\tag{10}$$

式中 $y(x)$——估测研究区域玉米的单产

x——特征变量的输入,即 VTCI 与 LAI

文献[26]基于 RF 算法确定了河北中部平原玉米各个生育时期的权重,进而构建了加权 VTCI 和 LAI 与玉米单产之间的双参数估产模型,结果显示,双参数估产模型精度相对较高,达到极显著的水平($P<0.001$),且基于随机森林算法确定的玉米各生育时期权重均较为合理。因此,本文引用其基于随机森林算法确定的河北中部平原玉米各生育时期的 VTCI 和 LAI 权重进行估产模型的构建。

1.3.3 组合预测模型

XGBoost 算法与 RF 算法均为统计理论的机器学习方法,但二者各有优势。XGBoost 算法通过引入正则化项在很大程度上避免了过拟合问题的出现,同时采用优化算法降低了问题计算复杂度,在处理大规模数据集方面具有明显的优势,已被广泛应用于不同领域的研究[27-29]。相对于 XGBoost,RF 更适合处理高维数据,对噪声数据的容忍性较高。通过结合 XGBoost 算法和 RF 算法的优点,参考经济学中合作博弈论的 Shapley 值理论,通过计算确定组合预测模型总误差在单一预测模型之间的分布,并在此基础上确定各单一预测模型的权重,具体步骤如下:

(1)求取 Shapley 值

$$E_i'=\sum_{\{s|i\in s\}}w(|s|)(E(s)-E(s-\{i\}))\tag{11}$$

其中 $$w(|s|)=\frac{(2-|s|)!\ (|s|-1)!}{2!}\tag{12}$$

式中 i——预测模型序号,即 XGBoost 与 RF 预测模型序号

E_i'——第 i 个预测模型的 Shapley 值,即 XGBoost 与 RF 预测模型计算得到的均方误差

s——包含模型 XGBoost 与 RF 的所有集合

$|s|$——预测模型个数

$E(s)$——集合 s 的收益

$E(s-\{i\})$——集合 s 中去除成员 i 后的收益

(2)计算权重

组合预测模型中若某单一预测模型所获得的误差分配值越大,表示预测精度越低,在组合预测模型中的权重就越小。基于此原则,预测模型 i 的权重 λ_i 定义为

$$\lambda_i=\frac{1}{n-1}\frac{E-E_i'}{E}\quad(i=1,2,\cdots,n)\tag{13}$$

最终的组合预测模型可描述为

$$Y=\sum_{i=1}^{n}\lambda_iY_i\tag{14}$$

式中 Y——组合预测模型估测的玉米单产

Y_i——模型 XGBoost 与模型 RF 估测的玉米单产

以通过计算得到的 2010—2017 年玉米 4 个生育时期的 VTCI 和 LAI 数据作为特征变量,相对应的玉米单产数据作为目标变量(每年 53 组数据,共 424 组数据),以 2010—2017 年(除 2012 年)数据作为训练集合,2012 年数据作为测试集合。由于特征变量与目标变量之间差值较大,特征变量值的范围

在 0 ~ 1 之间，而目标变量值在 3 000 ~ 9 000 kg/km^2 之间。由于特征变量和目标变量处于不同区间的值域差异可能对结果造成不同的影响，取对数后不会改变数据的性质和相对关系，对训练数据集和测试数据集中的目标变量进行取对数处理，并通过对估产模型输出的估测数据进行取对数还原，得到最终的估测单产数据。采用均方根误差（RMSE）、平均相对误差（MRE）和决定系数（R^2）等指标对模型的精度进行评价。

2　结果与分析

2.1　估产模型构建

构建 XGBoost 估产模型与 RF 估产模型，借鉴 Shapley 值理论，计算得到 XGBoost 与 RF 估产模型的 Shapley 值。根据 Shapley 值确定 XGBoost 估产模型与 RF 估产模型权重，进而完成对组合预测模型的构建。

2.1.1　XGBoost 与 RF 的估产模型构建

基于 XGBoost 与 RF 算法输出的特征重要性值，通过归一化计算得到研究区域玉米各生育时期 VTCI 与 LAI 的权重（表 1）。可以看出，玉米生育后期的 LAI 权重高于玉米生育前期 LAI 的权重，表明玉米生育前期的 LAI 对玉米产量影响程度较弱，玉米生育后期的 LAI 对玉米产量影响程度较强。原因可能是玉米 LAI 变化规律呈前期增长缓慢、中期增长快速、后期下降缓慢的偏峰曲线，其中玉米在出苗-拔节期和拔节-抽雄期主要以分化茎叶的营养生长为主，此阶段玉米生长迅速，叶片迅速增多增大；以抽雄-乳熟期为界，玉米进入以生殖生长为主的生育后期，光合产物的分配模式主要以果穗为中心，是玉米产量形成的重要时期。因此玉米生育后期的 LAI 对玉米产量的影响程度强于玉米生育前期。拔节-乳熟期的 VTCI 权重高于出苗-拔节期和乳熟-成熟期的权重，表明拔节-乳熟期的 VTCI 对玉米产量的影响程度较强，出苗-拔节期和乳熟-成熟期的 VTCI 对玉米产量的影响程度较弱。原因可能是在拔节-抽雄期、抽雄-乳熟期玉米进入营养生长的高峰期，此阶段玉米生长迅速，对土壤中的水分吸收也最为急迫，若发生水分胁迫将减缓玉米根茎叶的生长发育，降低光合作用对玉米干物质积累速率，减少蛋白质和有机质的合成，造成玉米粒重明显降低，最终影响玉米产量，因此拔节-抽雄期、抽雄-乳熟期的 VTCI 对玉米产量的影响程度较强。而在出苗-拔节期，玉米由于植株矮小，对水分的需求量较少，在乳熟-成熟期，玉米处于生育后期，生长变缓，对一定的水分胁迫表现出一定的忍受力，因此出苗-拔节期、乳熟-成熟期 VTCI 对玉米产量的影响程度较弱。综上所述，基于 XGBoost 算法与 RF 算法确定的研究区域玉米各生育时期的权重均较为合理。

表 1　玉米各生育时期的权重

Tab. 1　Weight results of each growth stage of maize

模型	LAI				VTCI			
	出苗-拔节期	拔节-抽雄期	抽雄-乳熟期	乳熟-成熟期	出苗-拔节期	拔节-抽雄期	抽雄-乳熟期	乳熟-成熟期
XGBoost	0.15	0.17	0.33	0.35	0.16	0.30	0.34	0.20
RF	0.18	0.12	0.27	0.43	0.19	0.25	0.34	0.22

分别将 XGBoost 估产模型与 RF 估产模型得到的玉米各生育时期的特征权重进行加权 VTCI 和 LAI 计算，进而构建基于加权 VTCI 和 LAI 与玉米单产之间的回归估测模型，将 2010—2017 年（除 2012 年）数据代入 XGBoost 估产模型与 RF 估产模型，进行可视化分析（图 2）。可以看出，XGBoost 估产模型 R^2 为 0.31，均方根误差为 940.91 kg/km^2，RF 估产模型 R^2 为 0.30，均方根误差为 947.50 kg/km^2，XGBoost 估产模型与 RF 估产模型均通过显著性检验（$P < 0.001$）。结果表明，XGBoost 估产模型与 RF 估产模型玉米单产估测精度均较高，可用于研究区域各县（区）玉米的单产估测。

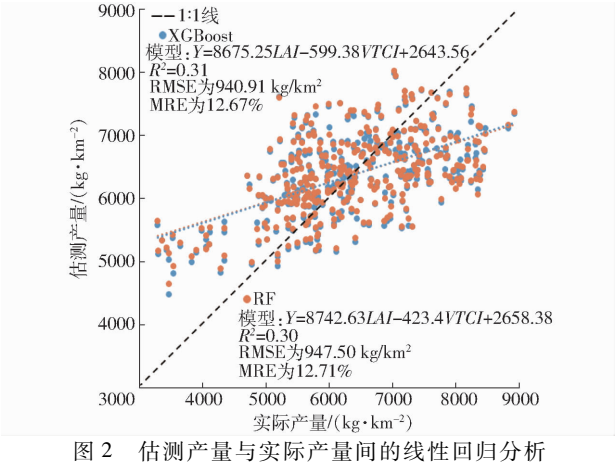

图 2　估测产量与实际产量间的线性回归分析

Fig. 2　Linear regression analysis between estimated and actual yields of maize

2.1.2　组合预测模型的构建

为进一步提高玉米单产估测精度，进行组合预测模型的构建。需要确定 XGBoost 与 RF 估产模型的权重，基于 2010—2017 年（除 2012 年）由单一估产模型输出的各县（区）的估测单产数据（共 371

组),分别计算 XGBoost 估产模型、RF 估产模型和组合预测模型的均方误差,将各模型的均方误差代入式(11),将得到的 XGBoost 与 RF 估产模型的 Shapley 值代入式(13),则可以得到 XGBoost 与 RF 估产模型的权重分别为 0.48 与 0.52,则组合预测模型为

$$Y = 0.48Y_{\mathrm{XGBoost}} + 0.52Y_{\mathrm{RF}} \tag{15}$$

通过计算,组合预测模型的 R^2 为 0.32,模型通过显著性检验且达显著水平($P < 0.001$)。同时,组合预测模型均方根误差为 939.81 kg/km^2,与单一 XGBoost 估产模型与 RF 估产模型相比,组合预测模型决定系数与均方根误差均得到提升。结果表明,组合模型的玉米单产估测精度优于单一估产模型,可用于河北中部平原的玉米单产估测。

2.2 玉米估测单产的精度评价

将 2012 年数据分别代入 XGBoost 估产模型、RF 估产模型与组合预测模型,将玉米的的估测单产与实际单产数据进行数据可视化分析(图 3)。对比分析,可以看出 XGBoost 估产模型、RF 估产模型和组合预测模型的估测单产与实际单产均呈显著的正相关关系。同时,R^2 均不小于 0.50,可以反映实际单产的波动均有 50% 以上能被估测单产的波动所描述,即玉米的实际单产与估测单产之间的误差较小。其中,组合预测模型 R^2 达到 0.52,均方根误差为 831.14 kg/km^2,平均相对误差为 9.86%,对比单一 XGBoost 估产模型与 RF 估产模型,组合预测模型的决定系数、均方根误差与平均相对误差均得到提升。进一步表明,基于 Shapley 值的组合预测模型的估产精度优于单一估产模型。

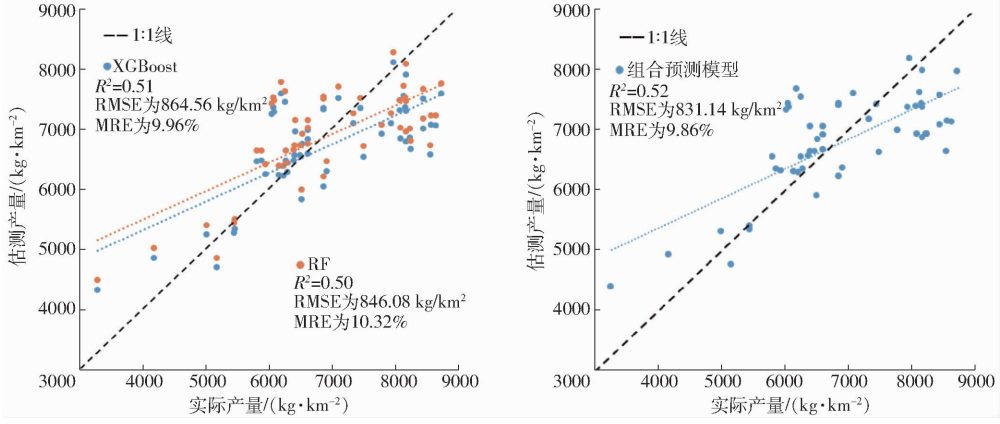

图 3 实际产量与估测产量散点图(2012 年)

Fig. 3 Scatter plots of estimated and actual yields of maize (2012)

2.3 区域玉米单产的估测与时空变化分析

将组合估产模型应用于河北中部平原 2010—2018 年逐像素玉米单产估测(图 4)。结果表明,研究区域玉米的估测单产随年际变化呈现先减少后增加的波动变化趋势。其中,2010—2014 年玉米估测单产整体呈现逐年下降的趋势,并在 2014 年玉米平均单产达到最低,平均单产在 6 000 kg/km^2 左右,原因可能是 2014 年河北中部平原降水较少且发生阶段性干旱,导致玉米单产减少严重;2010、2011、2013 年平均估测单产相差不大,均在 6 350 kg/km^2 左右。对比 2010—2014 年玉米估测单产年际变化趋势,2015—2018 年的玉米单产整体呈逐年上升的趋势,其中 2018 年玉米平均估测单产最高,约为 6 500 kg/km^2,2016 年和 2017 年玉米平均估测单产次之,约为 6 400 kg/km^2,2015 年玉米平均估测单产达到最低,约为 6 200 kg/km^2。

研究区域玉米的单产估测空间上呈现西部地区玉米估测单产最高,南部和北部地区玉米单产次之,东部地区玉米单产最低的分布特征。河北中部平原的北部地区中,2012 年玉米的估测单产最高,2014 年玉米估测单产最低,分别约为 6 000 kg/km^2 和 4 500 kg/km^2,2015、2016、2017 年玉米估测单产相差不大,均约为 5 000 kg/km^2,2011 年和 2018 年玉米的估测单产略低于 2012 年;东部地区中,2016 年玉米的估测单产最高,2014 年玉米的估测单产最低,分别约 5 000 kg/km^2 和 4 500 kg/km^2,2016—2018 年东部地区的玉米估测单产相差不大,均约为 4 700 kg/km^2。南部地区玉米的估测单产分别在 2012 年和 2014 年达到最高和最低,分别约 6 500、4 500 kg/km^2,其余年份玉米估测单产相差不大。西部地区中,2012、2015 年玉米的估测单产较高且相差不大,均约为 6 700 kg/km^2,2010、2014 年玉米估测单产较低,均约为 4 500 kg/km^2。

3 讨论

经过对不同估产模型特点的研究分析发现,在玉米单产估测过程中,不同估产模型对特征因素的信息提取方式不完全一样。同时以往单一估产模型

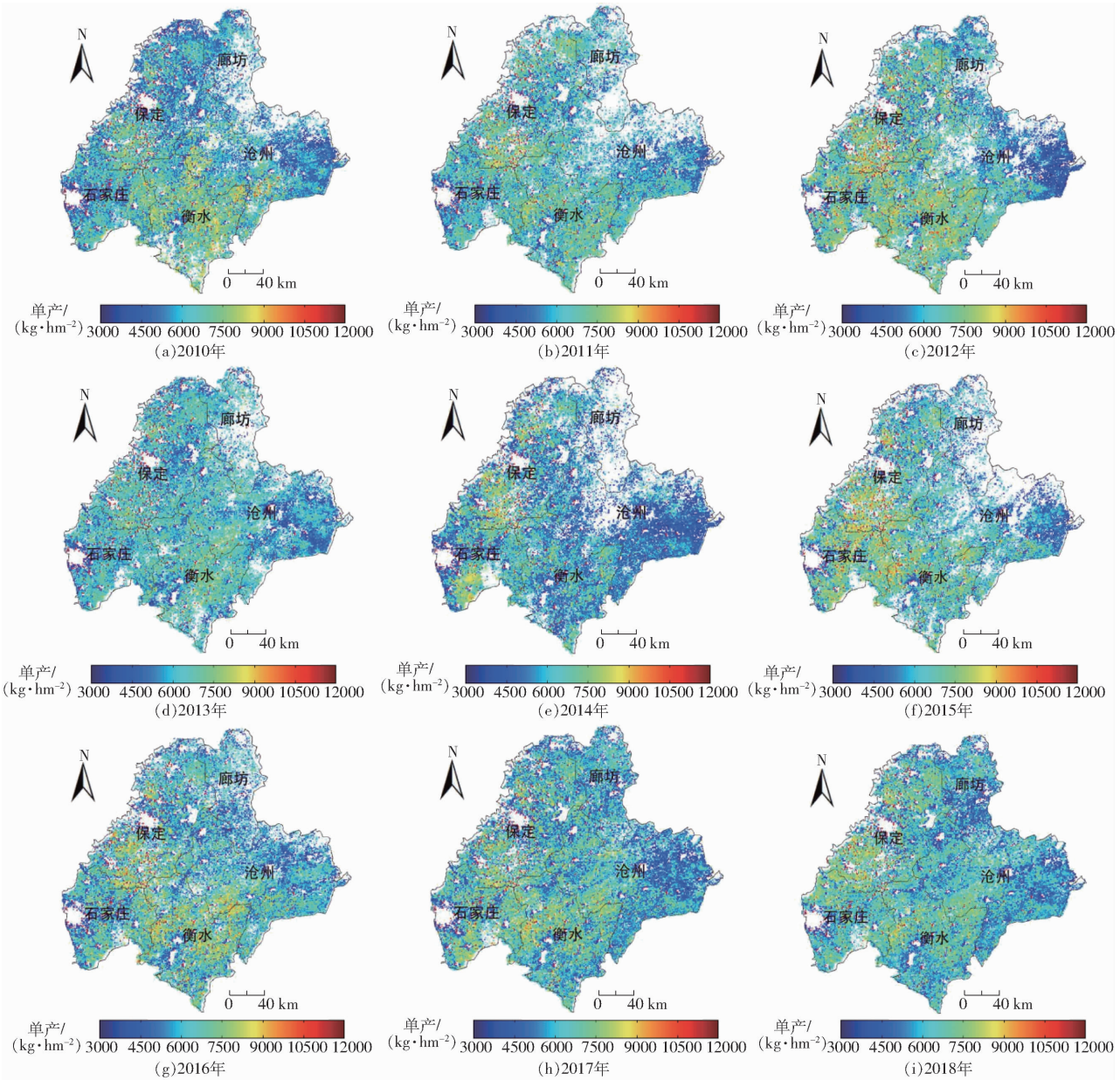

图 4　基于组合预测模型的玉米单产估测结果

Fig. 4　Estimated results of maize yield based on combination forecasting model

均注重特征因素对产量的影响，而缺少考虑估产模型对特征因素的信息提取以及对估测产量的影响。本文在考虑 VTCI 和 LAI 作物长势监测指标的基础上，也充分考虑了估产模型对特征因素的信息提取以及对估测产量的影响，利用 RF 对噪声数据的容忍性较高以及 XGBoost 较低的计算复杂度等优点，借鉴经济学中 Shapley 值理论计算得到 XGBoost 和 RF 估产模型的权重，进而构建组合估产模型，实现了单一估产模型的综合利用。在今后的研究中，可以在 Shapley 值理论的基础上尝试加入其他估产模型，以期进一步全面地提取特征因素的信息，避免单一估产模型在估产过程中造成有用信息的浪费，从而实现玉米单产估测精度的进一步提升。

本文虽然选取了与玉米长势和单产密切相关的 VTCI 和 LAI 作为特征变量，但玉米的生长发育除了受到水分胁迫和生长状态的影响之外，还受到其他自然因素和人为因素的影响，例如温度、土壤肥力、田间管理水平等因素。因此在基于 VTCI 与 LAI 作物生长指标的基础上，未来研究应进一步综合考虑与玉米单产相关性较大的其他因素。此外，基于 Shapley 值构建的组合预测模型对河北中部平原各县(区)玉米单产估测精度虽较高，但缺少对农学先验知识的考虑，今后研究中可以通过主观赋权法进一步修正 XGBoost 估产模型和 RF 估产模型权重，使模型权重更优。

4　结论

(1)通过借鉴组合预测模型思想，利用经济学

合作博弈论 Shapley 值利益分配理论确定了 XGBoost 估产模型与 RF 估产模型的权重,进而构建组合预测模型。结果表明,基于 Shapley 值的组合预测模型估测精度较高,且优于单一估产模型的精度。

(2)将组合预测模型应用于研究区域 2010—2018 年逐像素的玉米单产估测。结果表明,从研究区域玉米单产估测的年际变化看,河北中部平原 2010—2018 年各县(区)的玉米估测单产波动变化,但总体呈先减少后增加的趋势;从玉米单产估测的空间分布看,整体呈西部地区最高,南部和北部地区次之,东部地区最低的特征。

参 考 文 献

[1] 程国强,朱满德. 新冠肺炎疫情冲击粮食安全:趋势、影响与应对[J]. 中国农村经济,2020(5):13 - 20.
CHENG Guoqiang, ZHU Mande. COVID - 19 pandemic is affecting food security: trend, impacts and recommendations[J]. China Rural Economy, 2020(5): 13 - 20. (in Chinese)

[2] 顾晓鹤,何馨,郭伟,等. 基于 MODIS 与 TM 时序插补的省域尺度玉米遥感估产[J]. 农业工程学报,2010,26(增刊 2):53 - 58.
GU Xiaohe, HE Xin, GUO Wei, et al. Maize yield estimation at province scale by interpolation of TM and MODIS time-series images[J]. Transactions of the CSAE, 2010, 26(Supp. 2): 53 - 58. (in Chinese)

[3] 谭昌伟,罗明,杨昕,等. 运用 PLS 算法由 HJ - 1A/1B 遥感影像估测区域小麦实际单产[J]. 农业工程学报,2015,31(15):161 - 166.
TAN Changwei, LUO Ming, YANG Xin, et al. Remote sensing estimation of wheat practical yield on regional scale using partial least squares regression algorithm based on HJ - 1A/1B images[J]. Transactions of the CSAE, 2015, 31(15): 161 - 166. (in Chinese)

[4] SERVICE R F. The promise of drought-tolerant corn[J]. Science, 2009, 326(5952): 517.

[5] COOPER M, GHO C, LEAFGREN R, et al. Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product [J]. Journal of Experimental Botany, 2014, 65(21): 6191 - 6204.

[6] 王鹏新,龚健雅,李小文. 条件植被温度指数及其在干旱监测中的应用[J]. 武汉大学学报(信息科学版),2001,26(5):412 - 418.
WANG Pengxin, GONG Jianya, LI Xiaowen. Vegetation temperature condition index and its application for drought monitoring [J]. Geomatics and Information Science of Wuhan University, 2001, 26(5): 412 - 418. (in Chinese)

[7] 孙威,王鹏新,韩丽娟,等. 条件植被温度指数干旱监测方法的完善[J]. 农业工程学报,2006,22(2):22 - 26.
SUN Wei, WANG Pengxin, HAN Lijuan, et al. Further improvement of the approach to monitoring drought using vegetation and temperature condition indexes from multi-years' remotely sensed data[J]. Transactions of the CSAE, 2006, 22(2): 22 - 26. (in Chinese)

[8] WU B, MENG J, LI Q, et al. Remote sensing-based global crop monitoring: experiences with China's CropWatch system[J]. International Journal of Digital Earth, 2014, 7(2): 113 - 137.

[9] 林卉,梁亮,张连蓬,等. 基于支持向量机回归算法的小麦叶面积指数高光谱遥感反演[J]. 农业工程学报,2013,29(11):139 - 146.
LIN Hui, LIANG Liang, ZHANG Lianpeng, et al. Wheat leaf area index inversion with hyperspectral remote sensing based on support vector regression algorithm[J]. Transactions of the CSAE, 2013, 29(11): 139 - 146. (in Chinese)

[10] 史舟,梁宗正,杨媛媛,等. 农业遥感研究现状与展望[J/OL]. 农业机械学报,2015,46(2):247 - 260.
SHI Zhou, LIANG Zongzheng, YANG Yuanyuan, et al. Status and prospect of agricultural remote sensing [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2015, 46(2): 247 - 260. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 20150237&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2015.02.037. (in Chinese)

[11] 高辉明,张正斌,徐萍,等. 2001—2009 年中国北部冬小麦生育期和产量变化[J]. 中国农业科学,2013,46(11):2201 - 2210.
GAO Huiming, ZHANG Zhengbin, XU Ping, et al. Changes of winter wheat growth period and yield in Northern China from 2001—2009[J]. Scientia Agricultura Sinica, 2013, 46(11): 2201 - 2210. (in Chinese)

[12] 黎锐,李存军,徐新刚,等. 基于支持向量回归(SVR)和多时相遥感数据的冬小麦估产[J]. 农业工程学报,2009,25(7):114 - 117.
LI Rui, LI Cunjun, XU Xin'gang, et al. Winter wheat yield estimation based on support vector machine regression and multi-temporal remote sensing data[J]. Transactions of the CSAE, 2009, 25(7): 114 - 117. (in Chinese)

[13] 苏涛,王鹏新,刘翔舸,等. 基于熵值组合预测和多时相遥感的春玉米估产[J]. 农业机械学报,2011,42(1):186 - 192.
SU Tao, WANG Pengxin, LIU Xiangge, et al. Spring maize yield estimation based on combination of forecasting of entropy method and multi-temporal remotely sensed data[J]. Transactions of the Chinese Society for Agricultural Machinery, 2011, 42(1): 186 - 192. (in Chinese)

[14] TERUI N, DIJK H K V. Combined forecasts from linear and nonlinear time series models[J]. International Journal of Forecasting, 2002, 18(3): 421 - 438.

[15] BATES J M, GRANGER C W J. The combination of forecasts[J]. Operational Research Society, 1969, 20(4): 451 – 468.
[16] YUE Fang. Forecasting combination and encompassing tests[J]. International Journal of Forecasting, 2003, 19(1): 87 – 94.
[17] ANDRAWISA R R, ATIYAA A F, EL-SHISHINY H. Combination of long term and short term forecasts, with application to tourism demand forecasting[J]. International Journal of Forecasting, 2011, 27(3): 870 – 886.
[18] GRANGER C W J, RAMANATHAN R. Improved methods of combining forecasts[J]. Journal of Forecasting, 2010, 3(2): 197 – 204.
[19] 徐梅,晏福,刘振忠,等. 灰色 GM(1,1)–小波变换– GARCH 组合模型预测松花江流域水质[J]. 农业工程学报,2016,32(10):137 – 142.
XU Mei, YAN Fu, LIU Zhenzhong, et al. Forecasting of water quality using grey GM(1,1)– wavelet – GARCH hybrid method in Songhua River Basin[J]. Transactions of the CSAE, 2016, 32(10): 137 – 142. (in Chinese)
[20] 徐龙琴,张军,李道亮,等. 基于 EMD 和 ELM 的工厂化育苗水温组合预测模型[J/OL]. 农业机械学报,2016,47(4): 265 – 271,308.
XU Longqin, ZHANG Jun, LI Daoliang, et al. Combined prediction model of water temperature in industrialized cultivation based on empirical mode decomposition and extreme learning machine[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(4): 265 – 271,308. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20160435&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2016.04.035. (in Chinese)
[21] 薛昌颖,刘荣花,马志红. 黄淮海地区夏玉米干旱等级划分[J]. 农业工程学报,2014,30(16):147 – 156.
XUE Changying, LIU Ronghua, MA Zhihong. Drought grade classification of summer maize in Huang – Huai – Hai area[J]. Transactions of the CSAE, 2014, 30(16): 147 – 156. (in Chinese)
[22] 王鹏新,荀兰,李俐,等. 基于时间序列叶面积指数傅里叶变换的作物种植区域提取[J]. 农业工程学报,2017,33(21): 207 – 215.
WANG Pengxin, XUN Lan, LI Li, et al. Extraction of planting areas of main crops based on Fourier transformed characteristics of time series leaf area index products[J]. Transactions of the CSAE, 2017, 33(21): 207 – 215. (in Chinese)
[23] SUN W, WANG P, ZHANG S, et al. Using the vegetation temperature condition index for time series drought occurrence monitoring in the Guanzhong Plain, PR China[J]. International Journal of Remote Sensing, 2008, 29(17 – 18): 5133 – 5144.
[24] CHEN Tianqi, GUESTRIN C. XGBoost: a scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM, 2016.
[25] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001, 29(5): 1189 – 1232.
[26] 王鹏新,齐璇,李俐,等. 基于随机森林回归的玉米单产估测[J/OL]. 农业机械学报,2019,50(7):237 – 245.
WANG Pengxin, QI Xuan, LI Li, et al. Estimation of maize yield based on random forest regression[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(7): 237 – 245. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20190726&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2019.07.026. (in Chinese)
[27] 崔晓晖,师栋瑜,陈志泊,等. 基于 Spark 框架 XGBoost 的林业文本并行分类方法研究[J/OL]. 农业机械学报,2019,50(6):280 – 287.
CUI Xiaohui, SHI Dongyu, CHEN Zhibo, et al. Parallel forestry text classification technology based on XGBoost in Spark framework[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(6): 280 – 287. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20190632&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2019.06.032. (in Chinese)
[28] 齐建东,黄金泽,贾昕. 基于 XGBoost – ANN 的城市绿地净碳交换模拟与特征响应[J/OL]. 农业机械学报,2019,50(5): 269 – 278.
QI Jiandong, HUANG Jinze, JIA Xin. Simulation of NEE and charcterization of urban green-land ecosystem responses to climatic controls based on XGBoost – ANN[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(5): 269 – 278. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20190531&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2019.05.031. (in Chinese)
[29] 王鹏新,齐璇,李俐,等. 基于粒子群优化投影寻踪的玉米单产估测[J]. 农业工程学报,2019,35(13):145 – 153.
WANG Pengxin, QI Xuan, LI Li, et al. Estimation of maize yield based on projection pursuit with particle swarm optimization [J]. Transactions of the CSAE, 2019, 35(13): 145 – 153. (in Chinese)