基于大数据的小麦蚜虫发生程度决策树预测分类模型

张晴晴，刘勇，牟少敏，温孚江
山东农业大学农业大数据研究中心，山东 泰安 271018

摘要
小麦蚜虫是危害小麦的主要害虫。其发生程度预测特别是短期预测一直是植物保护领域难以解决的科学问题。传统预测方法通常仅采用温湿度，预测结果与实际发生匹配度不高。基于大数据的理念和数据挖掘技术，通过对2003-2013年小麦蚜虫发生程度与瓢虫、寄生蜂、日最高气压、日照时数等18种变量关系的决策树分析，构建了分类模型。经分析发现，日照时数与小麦蚜虫的发生程度关联度最高，其次是天敌瓢虫。该模型置信度为91.49%，且运行稳健。

关键词
小麦蚜虫；农业大数据；决策树；分类模型

中图分类号：S431.9 文献标识码：A
doi: 10.11959/j.issn.2096-0271.2016007

Decision tree predictive classification model on the occurrence degree of wheat aphids based on big data

ZHANG Qingqing, LIU Yong, MU Shaomin, WEN Fujiang
Agricultural Big Data Research Center, Shandong Agricultural University, Taian 271018, China

Abstract
Wheat aphids are the main pests of wheat crops. The monitoring and forecasting of their occurrence degree, especially the short-term occurrence degree, is much difficult. Many traditional predictions rely only on temperature and humidity, so the match degree to the actual occurrence value is low. Based on the concept of big data and data mining programs, the predictive classification model was established by means of the decision tree analysis of the relationship between the occurrence degree of aphids and up to 18 variables. It was found out that the duration of sunshine has the highest degree of relevance to the forecasting level of aphids, followed by ladybird. The confidence coefficient of the model that runs steadily in the experiment is 91.49%.

Key words
wheat aphids, agricultural big data, decision tree, classification model

1 引言

1.1 农业大数据研究背景

大数据的数据分析和挖掘功能可以发现数据间隐藏的相关性，并能通过合适的可视化技术将这种相关性形象地展示出来。这些都有助于人们揭示事物的发生和发展的基本规律，做出快速和符合实际的预测。2014年，荷兰科学家基于多年数据的深入分析，在《Nature》发文指出，新烟碱类杀虫剂吡虫啉的应用是本地食虫鸟类种群数量减少的主要原因[1]；同年，通过分析温度变化与埃塞俄比亚和哥伦比亚高海拔人口密集区的疟疾传播蚊虫和病原的数量变化关系，美国和英国科学家也在《Science》中发表论文，指明全球变暖会导致非洲和南美洲高原地带疟疾病的流行[2]。

农业大数据涉及农业领域的各个环节，采集、整合、挖掘和利用农业大数据，能够带来现代农业在农业生产、农业管理、农业经营和农业服务等方面的变革。农作物病虫害的监测预警是指导病虫害科学防控，保障国家粮食、食物和生态安全的重要前提。尽管目前病虫害监测预警已有一些专用的技术和软件，但往往采集数据指标偏少，数据挖掘技术不高，造成预警结果不准确，与实际发生匹配度不高；或者预测模型可操作性低，不能更好地直接为政府决策或农业生产服务。

1.2 小麦蚜虫的危害及监测预警

小麦是我国重要的粮食作物之一，山东省所处的黄淮海麦区是我国最主要的小麦产区。小麦长管蚜和禾谷缢管蚜是影响我国小麦生产的最主要害虫[3]。据统计，我国每年小麦蚜虫危害面积可达2.5亿亩，造成减产15%~30%，严重时可高达60%。近年来，全球气候变暖、耕作制度变化等因素使麦蚜的繁殖能力和适应性显著增强，其危害日趋严重[4]。

监测预警是小麦蚜虫“统防统治”和有效控制的基础。它能够严格按照防治阈值的要求服务于政府决策和农业生产。由于小麦蚜虫发生的普遍性和危害的严重性，国内外已有不少对其发生期和发生程度预测的研究[5-8]。但绝大多数研究仅仅是根据有限的气象数据（如温度和湿度），采用线性回归分析，建立相关的线性模型，开展中长期预测。此种预测忽视了生物因子（如自然天敌）及其他非生物因子与蚜虫发生的关联性，预测准确度低。因此，依据多年小麦蚜虫发生时农作物生长发育状况、气象条件、天敌因素、周边环境以及农事管理措施等数据的支撑，以大数据的研究技术，发挥其预测及分析功能，可为小麦蚜虫的绿色和科学防控服务。

1.3 决策树C5.0算法的发展及优势

决策树（decision tree）是一个类似于流程图的树结构，其中每个内部节点表示一个属性上的测试，每个分枝代表一个测试输出，而每个叶节点代表一种类别。决策树这一数据挖掘方法的起源是概念学习系统（concept learning system，CLS）。在CLS的基础上发展到ID3算法，ID3算法是该方法的高潮[9,10]。ID3算法是由Quinlan R于1986年提出的，他将Shannon的信息论引入决策树算法中，把信息熵作为选择测试变量的标准，对训练集进行分类并构造决策树来预测如何由变量对整个实例空间进行划分[11]，后来又演化为能处理连续变量的C4.5。最终C5.0算法出现，经过多

次改进该算法已经相对成熟，其主要优势体现在运行速度及性能方面。其另一优势是分析结果最终以树型图或者规则集的形式表示，不受时间的约束，将属性按照重要度权重大小排列在树型图上，在预测小麦蚜虫发生等级的过程中，可优先考虑重要度较大的属性，这样在判断麦蚜发生等级时较便捷，可充分满足实际生产的需求。

与SVM及神经网络只输出发生等级相比，C5.0树型图在田间的可操作性更强。决策树C5.0可通过人工干预，即决策树可以被修剪，避免模型的过度拟合。当然，常见的决策树算法有很多，如CHAID算法和CART算法，其中CHAID算法侧重于统计显著性检验；CART算法是根据Gini系数和方差来选择最佳分组变量和分割点，而C5.0算法以熵值函数将变量分组，在判断输入变量的异质性上，显然后者优于前者。因此，本文选用决策树C5.0作为构建小麦蚜虫发生程度模型的算法。

依据大数据理念，在农作物病虫害监测预警中，首次采用决策树的数据分析和挖掘手段，构建小麦蚜虫发生关联因子的决策树预测分类模型，为小麦蚜虫的有效控制，为保障国家粮食、食物安全和促进农业提质增效服务。

2 数据采集与分析

2.1 数据特征

本文涉及的数据类型主要包括2003-2013年小麦蚜虫的发生程度、天敌发生量、小麦生育期及逐日气象数据。其中，天敌有2类，分别是瓢虫和寄生蜂，瓢虫为平均1 m^2内的有效虫态数量，寄生蜂为平均百株僵蚜的数量；气象的变量种类共16个，分别为：20∶00-次日20∶00降水量、极大风速、极大风速的风向、平均本站气压、平均风速、平均气温、平均水汽压、平均相对湿度、日照时数、日最低本站气压、日最低气温、日最高本站气压、日最高气温、最大风速、最大风速的风向和最小相对湿度。变量中的2003-2013年小麦蚜虫的发生程度、天敌发生量和小麦生育期均来自鲁中生态区[12]各地植物保护站和本实验室逐年系统调查的数据。逐日气象数据来自国家气象中心。

2.2 数据预处理

2.2.1 目标变量离散化

在模型构建中，离散型变量要比连续型变量的处理速度快，因此将目标变量进行离散化处理[13]。根据中华人民共和国农业行业标准（NY/T612-2002）《小麦蚜虫测报调查规范》，当季蚜虫累计发生量达到发生总量的16%、50%、84%的时间分别为始盛期、高峰期、盛末期，从始盛期至盛末期一段时间为发生盛期。小麦蚜虫的发生程度分为5级，主要以当地小麦蚜虫发生盛期平均百株蚜量来确定，各级指标见表1。

2.2.2 变量删除

将可以用其他变量代替的变量删除，小麦生育期的变化基本与日照时数呈正相关，由于小麦生育期是通过观察小麦的生长发育情况人为确定的，其调查结果的误

表1 小麦蚜虫发生程度分级指标

级别	1级	2级	3级	4级	5级
百株蚜量（头，Y）	$Y\leqslant 500$	$500<Y\leqslant 1\,500$	$1\,500<Y\leqslant 2\,500$	$2\,500<Y\leqslant 3\,500$	$Y>3\,500$

差大于日照时数，因此小麦生育期保留日照时数。

2.2.3 决策树C5.0算法原理

决策树C5.0算法共涉及3个函数，分别是计算熵值函数、计算信息增益函数和计算信息增益率函数。其中，熵值函数是决策树的变量选择函数，用来预测信息位数。熵值函数的计算式如下：

$$\text{Entro}(p_1,p_2,\cdots,p_n)=-p_1\text{lb}(p_1)-\cdots-p_n\text{lb}(p_n) \quad (1)$$

其中，p_n为n发生的概率。Entro(p_1，p_2,…,p_n)=0，表示存在唯一的可能性；p_n的差别越小，Entro(p_1,p_2,…,p_n)的值就越大，相反，p_n的差别越大，熵值就越小。

决策树中信息熵的计算式如下：

$$\text{Info}(m)=-\sum_{i=1}^{k}((\text{freq}(n_i,m)/|m|)\times \text{lb}(\text{freq}(n_i,m)/|m|)) \quad (2)$$

其中，m是一个样本集合，目标变量n有k个，freq(n_i,m)表示n的样本数，|m|表示集合m的样本数。

根据计算所得的信息熵值计算信息增益值，信息增益函数是进行变量选择前后的信息差值的函数。S是某属性变量，有a个分类，其计算式如下：

$$\text{Info}(S)=-\sum_{i=1}^{a}((|S_i|/|S|)\times \text{Info}(S_i)) \quad (3)$$

Info(T)和Info(S)分别是决策树进行属性划分前后的信息增益值，其计算式如下：

$$\text{Gain}(S)=\text{Info}(T)-\text{Info}(S) \quad (4)$$

信息增益率则是逐个计算比较，最终确定树型图上节点的位置。

$$\text{GainRatio}(S)=\text{Gain}(S)/\text{Info}(S) \quad (5)$$

基于训练集得到的决策树通常不是最佳的，因为决策树中的构造会受到样本中异常数据的影响，存在过度拟合问题，得到的决策树因失去一般代表性而不适用于对新数据进行分类预测[14]。因此，需要对决策树进行剪枝。误差估计是在训练样本基础上给出一个置信度1-a，C5.0默认的置信度为1-25%=75%，然后计算错误率[15]。C5.0算法主要克服了ID3算法中偏向取值多的变量的不足[16]。本文采用IBM SPSS Modeler中较成熟的C5.0算法，最终结果可用树型图或者规则集的IF-THEN形式显示。

3 结果

将小麦蚜虫发生程度设置为目标变量，其余变量设置为输入变量，编写数据流。其中，样本中75%的数据作为训练集，25%的数据作为测试集，运行该数据流，得到信息增益率、树型图、规则集及准确率，建立相关的分类模型。

3.1 数据预处理及数据特性

经目标变量离散化及特殊值的去除，统计目标变量的结果如图1所示。其中，1级占最大比例，4级占比例最少，众数为1。

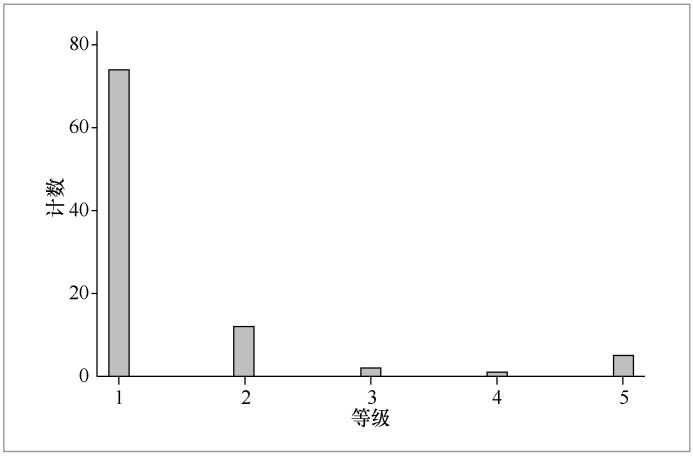


图1 小麦蚜虫发生等级统计

3.2 信息增益率

为消除训练集中的孤立点，决策树会对树型图进行剪枝训练，最终得到10个相关性较强的变量。其中，信息增益率最高的变量为日照时数（0.378 2），作为树型图的第一个节点进行测试。分别根据信息增益率的值分配各输入变量的节点位置，见表2。

表 2 输入变量的信息增益率

输入变量	信息增益率
最大风速（0.1 m/s）	0
平均水汽压（0.1 hPa）	0.004 3
日最低气温（0.1 ℃）	0.018 4
日最高本站气压（0.1 hPa）	0.051
20∶00–次日20∶00降水量（0.1 mm）	0.068 3
日最低本站气压（0.1 hPa）	0.080 1
极大风速风向（方位）	0.107
寄生蜂（头/百株）	0.130 2
瓢虫（头/m²）	0.162 5
日照时数（0.1 h）	0.378 2

3.3 决策树树型图

C5.0决策树的分析方法最终运行结果可用树型图的形式表示。图2为部分决策树树型图。

图2中，节点表示输入变量，其位置取决于信息增益率的大小。类别是目标变量的取值，即小麦蚜虫的发生等级，n表示样

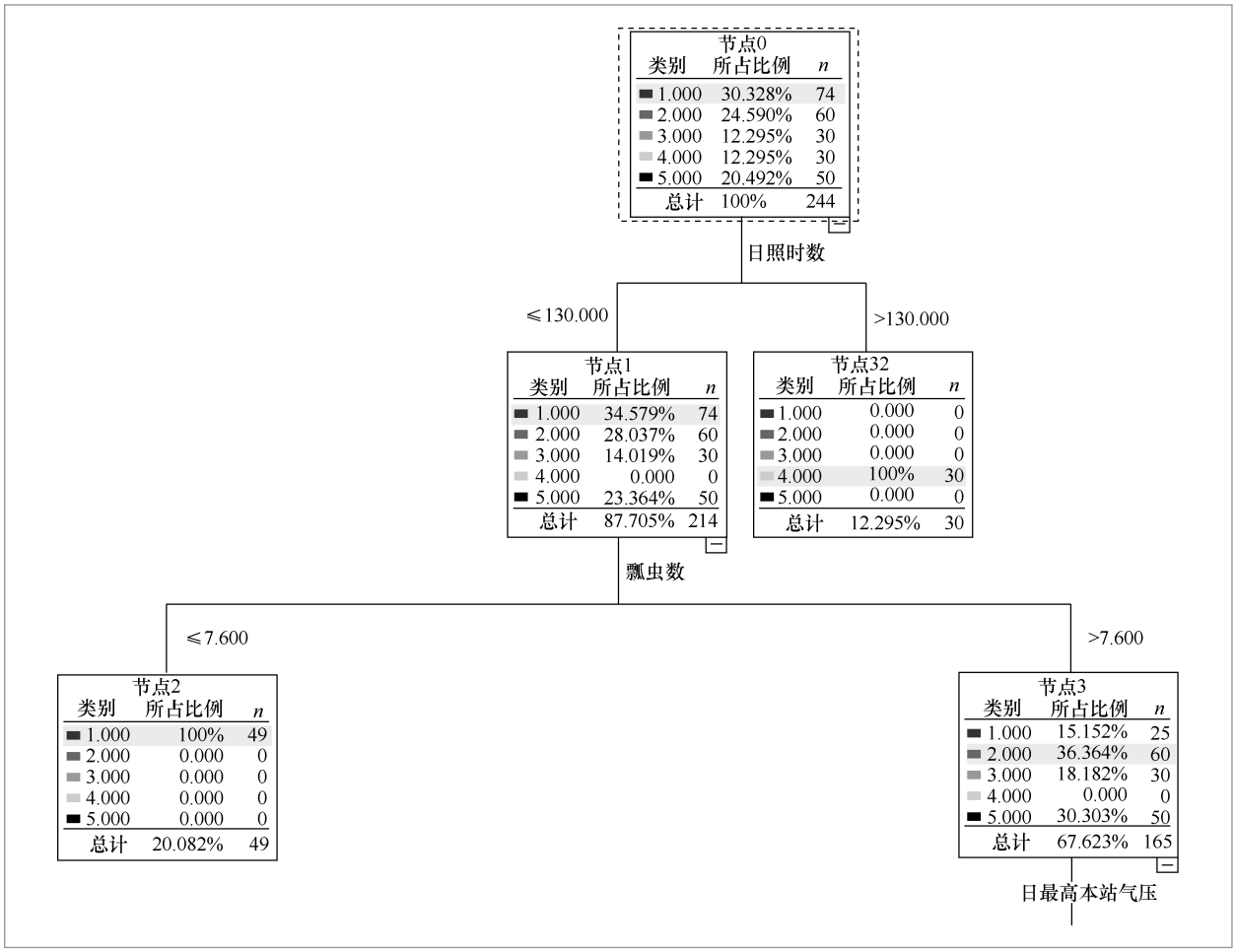


图 2 部分决策树树型图

本的个数。由图2可知，当日照时数大于13 h，小麦蚜虫发生程度为4级；当日照时数小于或等于13 h，并且百株瓢虫数小于或等于7.6头时，小麦蚜虫发生等级为1级；当百株瓢虫数大于7.6头时，如日最高本站气压大于848.9 hPa且20∶00-次日20∶00降水量大于0.3 mm，则小麦蚜虫发生等级为3级；当日最高本站气压大于848.9 hPa、20∶00-次日20∶00降水量小于或等于0.3 mm且最大风速大于11.5 m/s时，小麦蚜虫发生等级为2级，以此类推。

3.4 决策树规则集

规则集可根据树型图来提取，树型图中从头至尾的每一条执行线路为一条规则集，具体提取方法如图3所示。

决策树分析结果的另一种表达方式是“IF-THEN”的规则集形式。部分规则如下：

规则用于 1 - 包含 2 个规则

规则 1 用于 1.0

如果平均水汽压≤123

并且日照时数≤130

并且日最低气温>112

则 1.000

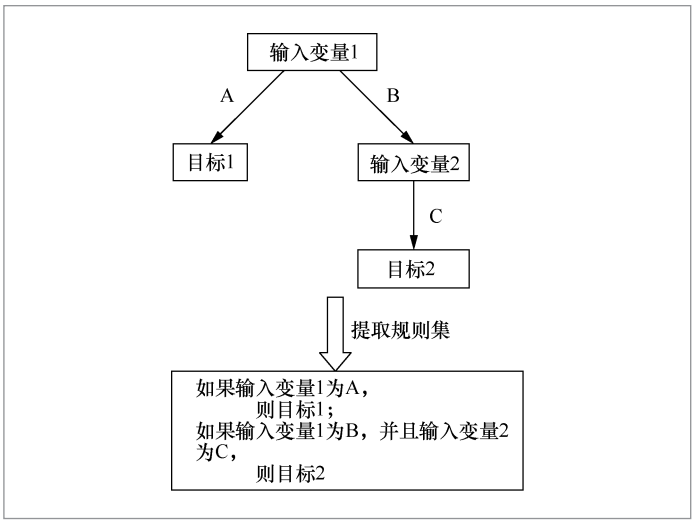


图 3 规则集提取流程

规则 2 用于 1.0

如果日照时数≤130

则 1.000

规则用于 3 - 包含 1 个规则

规则 1 用于 3.0

如果 20-20时降水量>3

并且日最高本站气压>8 489

则 3.000

规则用于 4 - 包含 1 个规则

规则 1 用于 4.0

如果日照时数>130

则 4.000

根据目标变量的取值，该规则集可分为5种规则，分别针对1级、2级、3级、4级和5级小麦蚜虫发生等级，由于原理类似，本文只呈现部分较短的规则。在每一种规则中包含一个或多个规则，可根据输入变量预测目标变量，无需再次进行计算机验证。例如规则用于1中的规则1，如果某年的平均水汽压小于或等于12.3 hPa，日照时数小于或者等于13 h，并且日最低气温大于11.2 ℃，那么小麦蚜虫的发生等级可预测为1级。

3.5 模型检验

C5.0算法常用于商业、医学等精确估计，其模型的置信度是统计预测值中正确值的个数占总样本数的比值。经分析，得到样本中目标变量的正确值与错误值。计算得到该模型的置信度为91.49%，且运算稳定（见表3）。由小麦蚜虫的预测值与真实值的拟合图可知（如图4所示），对于5级的预测效果偏离真实值最大。

4 讨论

农业大数据有其自身特有的复杂性和特殊性，相对于采用二维表来逻辑表达的

关系型数据结构，农业领域更多的是半结构化和非结构化数据，如大量的文字、图表、图片、动画语音、视频等形式组成的超媒体要素以及专家经验和知识农业模型等[17]，这些特性都使其更适合应用大数据技术。加之物联网技术向农业各领域渗透，大数据技术在农业上的应用已成为农业信息技术发展的必然趋势[18]。近年来，物联网技术在农业生产中的应用日渐深入，每年产生海量病虫害方面的数据，这些数据为农业大数据研究奠定了基础。大数据落脚于农业，让理论变为实践并服务于社会，引领现代农业进入新的发展空间，将会给农业带来翻天覆地的变化。

精准的数据挖掘并非依赖精准的算法，无论是分类模型还是回归模型，算法已经经历了无数次的运行验证，只有数据的质量才会决定最终结果的准确性。因此数据的预处理环节在数据挖掘中是非常重要的环节[19]。C5.0算法已经被验证无数次，其准确率高，主要是针对大数据集的分类算法，继续沿用C4.5算法的运算函数，运行速度和性能在C4.5的基础上有了明显提高。其结果最终呈现为非线性，无需假设输入变量间不相关。其优势在于分析结果为树型图或规则集的形式，在实际生产中无需运行算法，可直接辨别小麦蚜虫的发生等级。因此，根据本文研究结果，可采集气象、生育期及天敌参数，预测小麦蚜虫的发生等级，服务决策和农业生产。

表 3 小麦蚜虫预测模型的真实值与预测值

正确	错误	总计
223个	21个	244个
91.49%	8.51%	100%

小麦蚜虫的发生程度与气象因素和天敌的关联度高。本文淡化了调查的时间序列，随机选取训练集和测试集，通过训练集找出输入变量与目标变量之间的固定关系，然后用测试集验证这一关系。结果显示，小麦蚜虫的发生程度与日照时数关联度最高，其次为瓢虫和寄生蜂。因此，在小麦生产中，针对小麦蚜虫发生程度的短期预测，可依据该模型完成。另外，当日照时数达到13 h时，应当注意防控小麦蚜虫的大发生。

随着物联网数据采集技术在病虫害监测预警中的逐步应用，采集的规范化的海量数据会不断提高建模的数据质量，将会使预测更加符合实际。

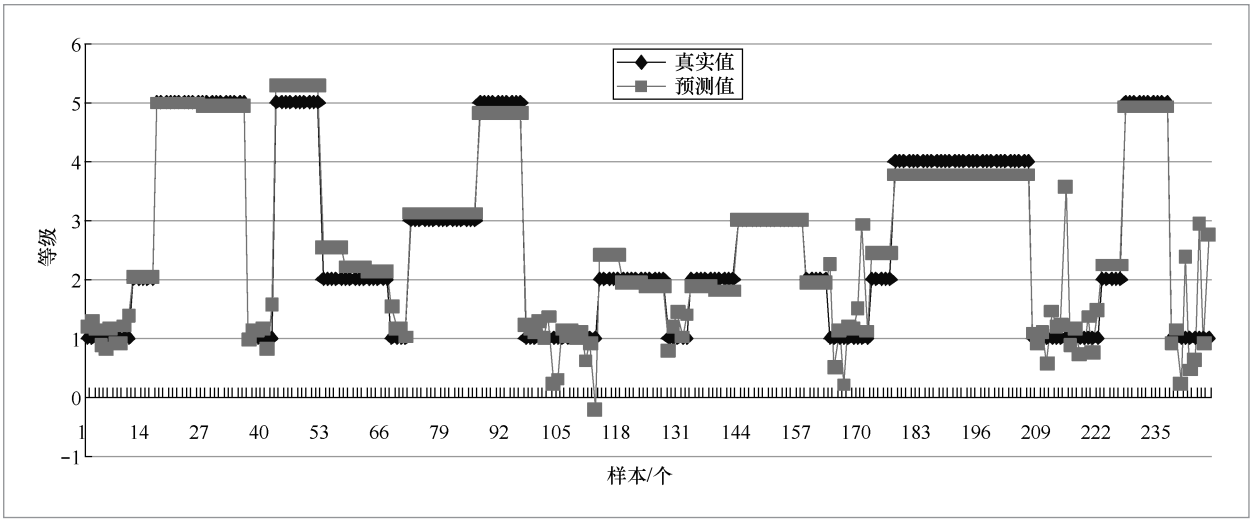


图 4 决策树分析结果拟合

参考文献:

[1] HHLLMANN C A, FOPPEN R P B, VAN TURNHOUT C A M, et al. Declines in insectivorous birds are associated with high neonicotinoid concentrations[J]. Nature, 2014, 511(7509): 341-343.

[2] SIRAJ A S, SANTOS-VEGA M, BOUMA M J. Altitudinal changes in malaria incidence in highlands of Ethiopia and Colombia[J]. Science, 2014, 343(6175): 1154-1158.

[3] 牟吉元. 农业昆虫学[M]. 北京: 中国农业科技出版社, 1995: 225-237.
MU J Y. Agricultural Entomology. Beijing: Chinese Agricultural Science and Technology Press, 1995: 225-237.

[4] 迟宝杰, 朱英菲, AXEL V, 等. 麦长管蚜及其天敌的种群发生和食物网分析[J]. 应用昆虫学报, 2014, 51(6): 1496-1503.
CHI B J, ZHU Y F, AXEL V, et al. Demographic and quantitative food web analysis of Sitobion avenae and its natural enemies[J]. Chinese Journal of Applied Entomology, 2014, 51(6): 1496-1503.

[5] PIYARATNE M K D K, ZHAO H Y, HU Z Q, et al. A model to analyze weather impact on aphid population dynamics: an application on swallow tall catastrophe model[J]. European Scientific Journal, 2014, 10(18): 1857-7431.

[6] DEBORAH J T, ART J D, FRANÇOISE A B, et al. Forecasting aphid outbreaks and epidemics of cucumber mosaic virus in lupin crops in a Mediterranean-type environment[J]. Virus Research, 2004, 100(1): 67-82.

[7] LUO J H, HUANG W J, ZHAO J L, et al. Predicting the probability of wheat aphid occurrence using satellite remote sensing and meteorological data[J]. Optik, 2014, 125(19): 5660-5665.

[8] 李文峰, 尹彬, 曹志伟, 等. 许昌市小麦蚜虫种群变化规律及气象预测模型[J]. 河南农业科学, 2011, 40(3): 81-84.
LI W F, YIN B, CAO Z W, et al. Variation of wheat aphid population in Xuchang and prediction models with meteorological data[J]. Journal of Henan Agricultural Sciences, 2011, 40(3): 81-84.

[9] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[10] QUINLAN J R. C4.5: Programs for Machine Learning[M]. Burlington: Morgan Kaufmanns Publishers, 1993: 69-81.

[11] 张家旺, 韩光胜, 张伟. C5. 0算法在RoboCup传球训练中的应用研究[J].计算机仿真, 2006, 23(4): 132-153.
ZHANG J W, HAN G S, ZHANG W. Application of C5.0 algorithmin passing ball training of RoboCup[J]. Computer Simulation, 2006, 23(4): 132-153.

[12] 于成. 基于cropwat的山东省主要粮食作物生产水足迹区域差异研究[D]. 济南: 山东师范大学, 2014: 9-11.
YU C. Study on regional of production water footprint of main crop based on cropwat in Shandong province[D]. Jinan: Shandong Normal University, 2014: 9-11.

[13] 朱廷勋, 高文. 基于数据挖掘的普通话韵律规则学习[J]. 计算机学报, 2000, 23(11): 1179-1184.
ZHU T X, GAO W. Data mining for learning mandarion prosodic models[J]. Journal of Computer Science, 2000, 23(11): 1179-1184.

[14] 刘军. 基于决策树算法的客户流失预测系统的分析与研究[D]. 武汉: 武汉理工大学, 2010: 45-54.
LIU J. Research of customer churn system based on decision tree algorithm[D]. Wuhan: Wuhan University of Technology, 2010: 45-54.

[15] 薛微, 陈欢歌. Clementine 数据挖掘方法及应用[M]. 北京: 电子工业出版社, 2010: 140-142
XUE W, CHEN H G. Clementine Data Mining Methods and Applications[M]. Beijing: Electronic Industry Press, 2010: 140-142.

[16] 陆安生, 陈永强, 屠浩文. 决策树C5算法的分析与应用[J]. 电脑知识与技术, 2005, 9(3): 17-20.
LU A S, CHEN Y Q, TU H W. The analysis and application of decision tree C5 algorithm[J]. Computer Knowledge and Technology, 2005, 9(3):17-20.

[17] 孙忠富, 杜克明, 郑飞翔, 等. 大数据在智慧农业中研究与应用展望[J]. 中国农业科技导报, 2013, 15(6): 63-71.

SUN Z F, DU K M, ZHENG F X, et al. Perspectives of research and application of big data on smart agriculture[J]. Journal of Agricultural Science and Technology, 2013, 15(6): 63-71.

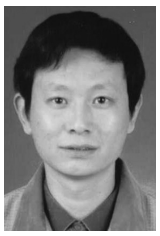
[18] 孙忠富, 杜克明, 尹首一. 物联网发展趋势与农业应用展望[J]. 农业网络信息, 2010(5): 5-8.
SUN Z F, DU K M, YIN S Y. Development trend of internet of things and perspective of its application in agriculture[J]. Agriculture Network Information, 2010(5): 5-8.

[19] 彭鸿涛, 聂磊. 发现数据之美——数据分析原理与实践[M]. 北京: 电子工业出版社, 2014: 5-7.
PENG H T, NIE L. Discover the Beauty of Data--Data Analysis Theory and Practice[M]. Beijing: Electronic Industry Press, 2014: 5-7.

作者简介

张晴晴(1991-), 女, 山东农业大学硕士生, 主要研究方向为农业大数据。

刘勇(1968-), 男, 山东农业大学教授、博士生导师, 主要研究方向为害虫绿色防控和农业大数据。



牟少敏(1964-), 男, 博士, 山东农业大学教授, 主要研究方向为大数据、机器学习和模式识别。

温孚江(1955-), 男, 现任山东农业大学校长、教授, 农业大数据创新战略联盟理事长, 全国人民代表大会常务委员会委员。早年留学美国, 并获博士学位。主要从事植物保护研究和宏观农业研究工作。发表论文210余篇, 专著5部。最近一部专著《大数据农业》由中国农业出版社于2015年9月出版。目前主要从事农业大数据应用研究工作, 是我国农业大数据研究主要发起人之一。

收稿日期: 2015-10-30
通信作者: 刘勇, liuyong@sdau.edu.cn; 温孚江, fjw@sdau.edu.cn
基金项目: 山东省农业重大应用技术创新课题基金资助项目
Foundation Item: Major Innovation of Applied Technology in Agriculture of Shandong Province