中国农业大学学报 2019,24(6):131-136
Journal of China Agricultural University
http://zgnydxxb.ijournals.cn
DOI:10.11841/j.issn.1007-4333.2019.06.15

基于机器学习的猪场用户流失分析

王 明[1] 滕光辉[1*] 宋维平[2]
(1.中国农业大学 水利与土木工程学院,北京 100083;
2.北京大北农科技集团股份有限公司,北京 100080)

摘 要 为实现养殖互联网平台用户的自动化运营,防止用户流失,采用区间估计和机器学习建模方法,对猪场用户的平台使用情况进行用户流失分析。结果表明:1)猪场规模越小,用户流失的概率越大,其中 500<母猪数 x≤1 000 的猪场规模用户流失的概率最小,属于忠诚用户;2)使用决策树算法、kNN 算法、贝叶斯分类算法进行猪场用户流失分析建模,3 种分类算法中,从平均识别率(F_1 度量)考虑,决策树模型的平均识别率(0.93)高于 kNN 模型(0.91)和贝叶斯分类模型(0.80),选择决策树算法用于平台用户流失分析建模是可行的。本研究所构建的猪场用户流失分析模型,可为养殖互联网平台的猪场用户研究提供可靠的数据模型,实现平台用户的自动化运营,也可为其他类型的用户研究和产品设计提供参考。

关键词 猪场;互联网+农业;用户流失;机器学习;决策树
中图分类号 X24 **文章编号** 1007-4333(2019)06-0131-06 **文献标志码** A

Churn analysis of pig farm users based on machine learning

WANG Ming[1], TENG Guanghui[1*], SONG Weiping[2]
(1. College of the Water Resources and Civil Engineering, China Agricultural University, Beijing 100083, China;
2. Beijing Dabeinong Technology Group Co. Ltd., Beijing 100080, China)

Abstract In order to realize the automatic operation of users and prevent user churn, interval estimation and machine learning modeling methods are adopted to analyze user churn by data records of pig farm users. The results show that: 1) The smaller the pig farm is, the greater the possibility of user churn is. The least possible churn of pig farm users, whose number of sows x is at 500 < x ≤ 1 000, belong to loyal users; 2) Comparing three different classification algorithms, e.g. the decision tree algorithm, kNN algorithm, Bayesian classification algorithm in pig farm users churn modeling, the average recognition rate of decision tree model (0.93) is higher than the kNN model (0.91) and Bayesian classification model (0.80). Therefore, choose the decision tree algorithm for platform user churn analysis is feasible. The research indicates that the churn analysis of pig farm users model could provide customized services for users of different sizes, provide a reliable data model for user research of other categories and realize the automatic operation of platform users.

Keywords pig farm; internet + agriculture; users churn; machine learning; decision tree

“互联网+农业”是一种生产方式、经济模式与技术手段的创新,推动了互联网技术在养殖业的快速发展,而基于互联网市场的便捷性与开放性,在一定程度上加剧了市场竞争,使得用户容易在不同平台间转换,从而产生用户流失[1-2]。随着流量红利消失,用户和市场占有率增速下降。设计智能化的互联网产品,实现用户的自动化运营,防止用户流失具有重要的现实意义[3-4]。

养殖业互联网平台必须从流量运营向用户运营转变,用户运营实质是数据化运营或精准化运营,而

收稿日期:2018-09-18
基金项目:国家重点研发计划项目(2016YFD0700204)
第一作者:王明,博士研究生,E-mail:wangmingwang2008@163.com
通讯作者:滕光辉,教授,主要从事农业智能装备、数字农业等研究,E-mail:futong@cau.edu.cn

流量运营与用户运营相比，更像是粗放式运营。已有研究对用户流失问题提出了一系列理论以及科学的分析方法：1）用户运营依赖于会员生命周期理论、会员 RFM 理论、会员增长理论、会员终身价值理论以及会员权益体系等[5-6]；2）流量运营在这些方面主要基于以前的经验或线下经验，缺乏系统规划[7]；3）采用区间估计和机器学习建模对互联网平台用户流失进行分析预测，这样可以采取相应措施来挽留即将流失的客户，以此减少利润损失[8-9]。4）新用户开发所用的成本是老用户维持所用成本的 4～5 倍[10-11]。其中，采用区间估计和机器学习建模对互联网平台用户流失进行分析预测是比较可行的研究方法[12-13]。由于养殖业是传统产业，亟需借助互联网技术提高产业效率。然而，基于养殖业互联网平台的用户研究理论存在不少薄弱环节。

鉴于此，本研究拟采用区间估计和机器学习建模方法对猪场用户的平台使用情况进行分析，构建猪场用户流失分析模型，以期为养殖业互联网平台用户流失研究提供产品设计依据和数据模型。

1 材料与方法

1.1 研究对象及方法

本试验使用的数据由国内某农业互联网平台运营商提供，猪场用户数据包含 151 个 Excel 文件，时间跨度为 27 个月，共计 260 956 条数据记录，具体包括公猪、母猪、后备种猪、育肥猪、单据合计等属性信息，主要指业务模块在农场生产管理和流通交易过程中产生的单据信息。

本研究首先采用区间估计分析用户流失的影响因素，其次采用机器学习方法来构建猪场用户流失分析模型，衡量互联网平台用户流失的程度，最后验证评估建模效果。具体包括，结合猪场生产管理和流通交易的业务场景，通过数据调和，将基础数据转化为数据模型所需的数据形式，进行用户流失分析建模，并针对模型进行评估。用户流失分析问题是一个不平衡的分类问题，使用错误率和精度往往难以判断模型的好坏，准确率、召回率、F_1 值、“受试者工作特征”曲线（简称 ROC 曲线）是更为适用于此类问题的性能度量[14]。

1.2 数据预处理

对各个属性进行数据预处理，可以得到其数据量、均值、标准差、最小值、下四分位数、中位数、上四分位数、最大值等信息。判断异常值的数据规则有奈尔检验法和格拉布斯检验法。如果标准差已知，采用奈尔检验法；如果标准差未知，采用格拉布斯检验法[15]。本研究发现，公猪、肥猪、后备、销售额等数据属性存在离群点。为保证数据的可靠性与准确性，进行数据清洗，去除不完整数据，并对离群点做清理工作。

2 结果与分析

2.1 用户流失的影响因素

养殖业互联网平台用户流失的影响因素很多，包括猪场规模大小、单据合计、产品商业变现、市场补贴推广、同行竞争等，由于缺乏产品商业变现、市场补贴推广、同行竞争等方面的数据支撑，本研究主要分析猪场规模因素和单据合计因素，并通过区间估计方法来量化该因素，推断不同规模大小猪场用户流失的概率。

猪场用户数据量较大，假设符合正态分布。将猪场用户数据按照猪场规模大小进行特征提取，以猪场母猪数（x）进行猪场规模分组，依次为：$x\leqslant 50$，$50<x\leqslant 200$，$200<x\leqslant 500$，$500<x\leqslant 1\,000$，$x>1\,000$。$x\leqslant 50$ 的猪场规模用户的单据合计数均值为 48，标准差为 88.8，呈正偏（右偏）分布，这可以通过对平均值和中值作比较进行验证，中值小于平均值，进而确定呈正偏（右偏）分布。应用对数转换对偏态分布进行修复，图 1 示出 $x\leqslant 50$ 的猪场规模用户的单据合计数直方图，单据合计数均值为 3.1，标准为 1.3，其分布可拟合为正态分布。

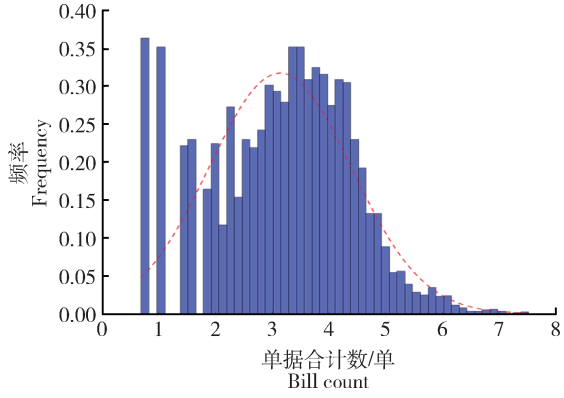
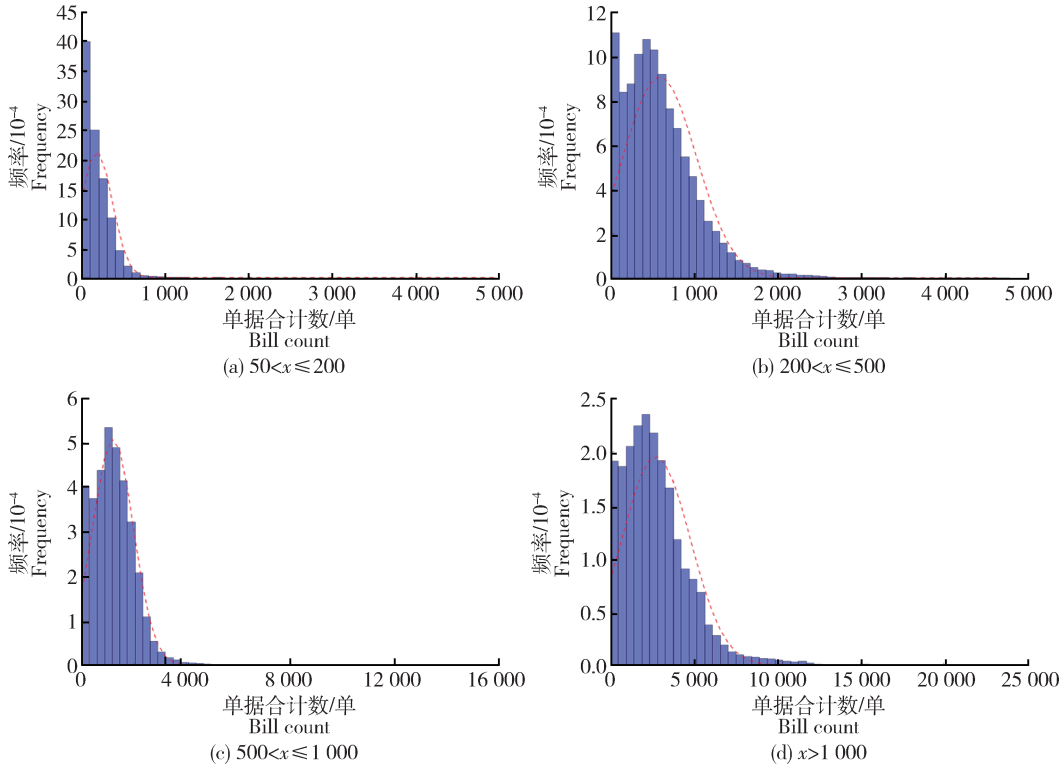


图 1 母猪数 $x\leqslant 50$ 猪场规模用户的单据合计数直方图

Fig. 1 Bill count histogram for pig farm users whose number of sows is no more than 50

图 2 示出不同猪场规模用户的单据合计数直方图：$50<x\leqslant 200$ 猪场规模用户的单据合计数的均值为 179.6，标准差 185.6；$200<x\leqslant 500$ 猪场规模用

户的单据合计数的均值为 581.0，标准差 438.9；500＜x≤1 000 猪场规模用户的单据合计数均值为 1 213.6，标准差 784.5；母猪数 x＞1 000 猪场规模用户的单据合计数的均值为 2 751.1，标准差为 2 046.7。通过上述分析，不同规模猪场用户的单据合计数均呈正态分布，定义单据合计数＜50 的用户有流失的可能，因此可用 Z 值计算不同规模猪场用户流失的概率。



x 为母猪数。The number of sows is represented by x.

图 2　不同猪场规模用户的单据合计数直方图

Fig. 2　Bill count histogram of pig farm users of different sizes

Z 值采用式(1)计算：

$$Z = (y - \mu)/\sigma \tag{1}$$

式中：y 为单据合计数，单；μ 为总体均值，单；σ 为总体标准差，单。

对不同规模猪场用户流失的概率进行区间估计，x≤50 的猪场规模用户流失的概率为 73.6%，50＜x≤200 的猪场规模用户流失的概率为 24.2%，200＜x≤500 的猪场规模用户流失的概率为 11.5%，500＜x≤1 000 的猪场规模用户流失的概率为 6.9%，x＞1 000 的猪场规模用户流失的概率为 9.5%。猪场规模越小，用户流失的概率越大，其中 500＜x≤1 000 的猪场规模用户流失的概率最小，属于忠诚用户。

2.2　猪场用户流失分析模型的构建

研究用户流失时，通过机器学习方法来衡量用户流失的程度。使用表征猪场规模的公猪、母猪、后备种猪和育肥猪这 4 项字段作为数据集，使用表征平台使用情况的单据合计字段作为标签集。依次将每个 Excel 文件中的数据导入到内存中，并进行数据清洗，最终得到数据集 X 和标签集 Y。其中数据集 X 中为公猪、母猪、后备种猪和育肥猪，标签集 Y 中为是否为活跃用户。经过大量观察和分析，定义单据合计≥50 的猪场为活跃用户，单据合计数＜50 的猪场为非活跃用户，意味着用户离开或流失。在技术处理上，将活跃用户标记为 1，非活跃用户标记为 0。

一般而言，逻辑回归按因变量的资料类型分为二分类和多分类 2 种，其中二分类的应用频率更高，也更加容易理解，本研究的猪场用户流失分析模型也属于二分类的逻辑回归范畴。尝试进行逻辑回归分析，逻辑回归模型拟合需要以下步骤：1)根据机器学习需求设计指标变量；2)列出逻辑回归方程；3)估

计回归系数;4)应用方差分析表对回归模型进行 F 检验;5)应用参数估计表对回归系数进行 t 检验;6)模型应用:控制自变量的取值,输入模型后得到预测变量的值[16]。最后,逻辑回归模型的平均正确率为0.83。

从猪场用户数据中取80%的数据用于构建猪场用户流失分析模型,共计206 618条记录。猪场用户数据的系列特征值可作为模型训练和评测的数据,研究中分别采用决策树算法、kNN算法、贝叶斯分类算法来构建模型,探寻分析猪场规模与用户流失的关系。

首先,在筛选用户流失建模因子的时候,需要重点考虑数据集与标签集的之间的关系以及解释度,本研究选取了公猪、母猪、后备种猪和育肥猪这四个变量作为搭建用户流失分析模型的因子。其次,提取活跃用户数据和非活跃用户数据,进行数据导入、数据描述、去除建模中不需要的字段,开始构建决策树模型,描述模型(变量显著)。最后,验证决策树模型。kNN模型和贝叶斯分类模型的构建过程与之类似。

1)决策树是基于树结构来进行决策的,这恰是人类在面临决策问题时一种很自然的处理机制。一般的,一棵决策树包含一个根结点、若干个内部结点和若干个叶结点;叶结点对应于决策结果,其他每个结点则对应于一个属性测试;每个结点包含的样本集合根据属性测试的结果被划分到子结点中;根结点包含样本全集。从根结点到每个子结点的路径对应了一个判定测试序列。决策树学习的目的是为了产生一棵泛化能力强,即处理未知示例能力强的决策树,其基本流程遵循简单且直观的"分而治之"策略[17-18]。

2)k 近邻(简称kNN算法)监督学习的算法流程:给定测试样本,基于某种距离度量找出训练集中与其最靠近的 k 个训练样本,然后基于这 k 个"邻居"的信息来进行预测。通常,在分类任务中选择这 k 个样本中出现最多的类别标记作为预测结果;还可基于距离远近进行加权平均或加权投票,距离越近的样本权重越大。kNN在训练阶段仅仅是把样本保存起来,待收到测试样本后再进行处理[19]。假设有 N 种可能的类标记,即 $y=\{c_1,c_2,\cdots,c_N\}$,给定测试样本 x,若其最近邻样本为 z,$P(c|x)$ 和 $P(c|z)$ 均为后验概率,则最近邻分类器出错的概率 P_e 即为样本 x 与样本 z 的类标记不同的概率,即

$$P_e = 1 - \sum_{c \in y} P(c|x)P(c|z) \quad (2)$$

3)贝叶斯学习是概率框架下实施决策的基本方法。对分类任务来说,在所有相关概率都已知的理想情形下,贝叶斯学习考虑如何基于这些概率和误判损失来选择最优的类标记[20-22]。假设有 N 种可能的类标记,即 $y=\{c_1,c_2,\cdots,c_N\}$,对每个样本 x,选择能使后验概率 $P(c|x)$ 最大的类标记,则最小化分类错误率的贝叶斯最优分类器 $h^*(x)$ 为

$$h^*(x) = \mathrm{argmax}_{c \in y} P(c|x) \quad (3)$$

2.3 模型验证评估

从猪场用户数据中取剩余的20%数据用于验证评估数据模型,共计51 655条记录。基于机器学习模型评估的一般框架,分别对上述的决策树算法、kNN算法、贝叶斯分类算法进行建模验证,得出了非活跃用户和活跃用户分类识别的预测准确率、召回率、F_1 度量和实际值。其中,预测准确率指正确预测个数占总预测个数的比例,召回率指正确预测个数占实际个数的比例,F_1 度量为召回率和预测准确率的调和平均值。决策树模型的分类识别结果见表1,kNN模型和贝叶斯分类模型的分类识别结果与之类似。3种分类算法中,单从平均识别率(F_1 度量)考虑,决策树模型的平均识别率(0.93)高于kNN模型(0.91)和贝叶斯分类模型(0.80),这充分体现了这种基于属性特征阈值定义分类规则和停止规则的决策树算法在用户流失分析中的优越性,而

表1 决策树模型的分类识别结果

Table 2 Classification identification result of decision tree model

用户类别 User class	预测准确率 Precision	召回率 Recall	F_1 度量 F_1-score
非活跃用户 Inactive user	0.78	0.79	0.79
活跃用户 Active user	0.96	0.95	0.96
加权平均 Weighted average	0.93	0.93	0.93

kNN 算法和贝叶斯分类算法识别运算需要依赖于训练过程中的样本质量和识别类型的划分。由此可知，选择决策树算法用于平台用户流失分析建模是可行的。

机器学习分类模型是为测试样本产生一个实值或概率预测，然后将这个预测值与一个分类阈值进行比较，若大于阈值则分为正类，否则为反类。根据这个实值或概率预测结果，可将测试样本进行排序，“最可能”是正例的排在最前面，“最不可能”是正例的排在最后面，排序本身的质量好坏，体现了学习器在不同任务下的“期望泛化性能”的好坏，ROC 曲线就是从这个角度出发来研究学习器泛化性能的有力工具[23-25]。决策树模型的 ROC 曲线见图 3，表示在不同分类阈值下，真阳性率和假阳性率的关系，kNN 模型和贝叶斯分类模型进行分类时的 ROC 曲线与之类似。ROC 曲线越靠近左上角、与横轴围成的面积越大，代表该模型对于数据的分类预测效果越好，与 kNN 模型和贝叶斯分类模型相比，决策树模型的 ROC 曲线与横轴围成的面积较大。综上，选择决策树算法用于平台用户流失分析建模是可靠的。

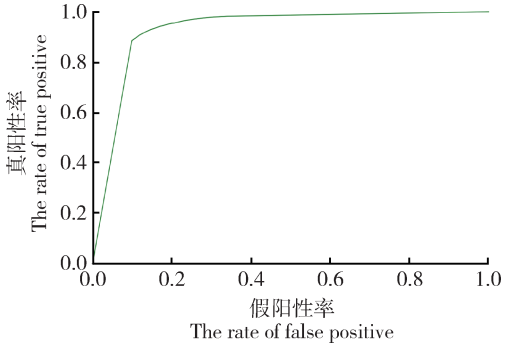


图 3　决策树模型的 ROC 曲线

Fig. 3　ROC curve of decision tree model

3　结　论

本研究采用区间估计和机器学习方法对猪场用户的平台使用情况进行分析，构建了猪场用户流失分析模型。主要结论如下：

使用区间估计方法进行用户流失的影响因素分析，掌握不同规模的猪场用户流失概率的大小，有助于养殖业互联网平台的产品设计，针对不同类型的用户提供定制化的服务，增强用户粘性，防止用户流失。

使用决策树算法、kNN 算法、贝叶斯分类算法进行猪场用户流失分析建模，3 种分类算法中，从平均识别率（F_1 度量）考虑，决策树模型的平均识别率（0.93）高于 kNN 模型（0.91）和贝叶斯分类模型（0.80），选择决策树算法构建猪场用户流失分析模型是可行的，可为猪场用户和其他类型用户研究提供可靠的数据模型，实现平台用户的自动化运营。

参考文献　References

[1] 周丽莎，李娟. 移动互联时代的用户体验[J]. 通信企业管理，2013(4)：18-19
Zhou L S，Li J. User experience in the mobile internet era[J]. *C-Enterprise Management*，2013(4)：18-19 (in Chinese)

[2] 李海舰，田跃新，李文杰. 互联网思维与传统企业再造[J]. 中国工业经济，2014(10)：135-146
Li H J，Tian Y X，Li W J. Mobile internet thinking and traditional business reengineering[J]. *China Industrial Economics*，2014(10)：135-146 (in Chinese)

[3] 郭顺利，张向先，相甍甍. 高校图书馆微信公众平台用户流失行为模型及其影响因素分析[J]. 图书情报工作，2017，61(2)：57-66
Guo S L，Zhang X X，Xiang M M. Research on the customer churn behavior model and its influencing factors of WeChat public platform in university libraries[J]. *Library and Information Service*，2017，61 (2)：57-66 (in Chinese)

[4] 谭莹. 我国生猪生产效率及补贴政策评价[J]. 华南农业大学学报：社会科学版，2010，9(3)：84-90
Tan Y. Analysis on live pig production efficiency and live pig subsidy policy in China[J]. *Journal of South China Agricultural University：Social Science Edition*，2010，9 (3)：84-90 (in Chinese)

[5] 徐孝娟，赵宇翔，朱庆华. 社交网站用户流失行为理论基础及影响因素探究[J]. 图书情报工作，2016，60(4)：134-141
Xu X J，Zhao Y X，Zhu Q H. Theoretical basis and influence factors of user exodus behavior of social networking sites[J]. *Library and Information Service*，2016，60 (4)：134-141 (in Chinese)

[6] 刘勇，万维新，康子轶. 中国电信业流失客户分类研究[J]. 统计教育，2009(1)：26-30
Liu Y，Wan W X，Kang Z Y. Study on the lost customers classification of Chinese telecom industry[J]. *Statistical Education*，2009(1)：26-30 (in Chinese)

[7] 江积海，张烁亮. 平台型商业模式创新中价值创造的属性动因及其作用机理[J]. 中国科技论坛，2015(7)：154-160
Jiang J H，Zhang S L. On the internal attributes and mechanism in value creation of platform-based business model innovation[J]. *Forum on Science and Technology in China*，2015(7)：154-160 (in Chinese)

[8] 徐孝娟,赵宇翔,朱庆华,孙霄凌.社交网站中用户流失要素的理论探讨及实证分析[J].信息系统学报,2013(2):83-97
Xu X J,Zhao Y X,Zhu Q H,Sun X L. Theoretical discussion and empirical analysis of user loss factors in social networking sites[J]. *China Journal of Information Systems*,2013(2):83-97 (in Chinese)

[9] 潘红英,郑卫兵,王德刚.猪群死亡分布规律分析及相关控制技术探讨[J].浙江农业科学,2005,46(3):221-223
Pan H Y,Zheng W B,Wang D G. Analysis of distribution pattern of pig death and preventing approaches[J]. *Journal of Zhejiang Agricultural Sciences*, 2005, 46(3): 221-223 (in Chinese)

[10] 李倩,钟胜.面向管理改进的服务企业顾客满意度模型[J].商业经济与管理,2005(4):66-71
Li Q, Zhong S. The customer satisfaction model for management improvement of service enterprises[J]. *Journal of Business Economics*,2005(4):66-71 (in Chinese)

[11] 王永贵,董大海.客户关系管理的研究现状、不足和未来展望[J].中国流通经济,2004,18(6):52-56
Wang Y G,Dong D H. Customer relationship management: Current situation,weakness and future research directions[J]. *China Business and Market*,2004,18(6):52-56 (in Chinese)

[12] 严浩仁,贾生华.顾客忠诚的基本驱动模型研究:以移动通信服务为例[J].经济管理,2005(4):42-46
Yan H R,Jia S H. Analysis on the basic drive model of customer loyalty: The case study of mobile communication service[J]. *Business Management Journal*,2005(4):42-46 (in Chinese)

[13] 李震.互联网平台如何创造体验价值:基于互动视角的分析[J].广东财经大学学报,2017,32(2):15-30
Li Z. How to create experiential value based on internet platform:An analysis from the perspective of interaction[J]. *Journal of Guangdong University of Finance & Economics*, 2017,32(2):15-30 (in Chinese)

[14] Spackman K A. Signal detection theory: Valuable tools for evaluating inductive learning[C]. In: *Proceedings of the sixth International Workshop on Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.,1989:160-163

[15] 卢元磊,何佳洲,安瑾,苗高洁.几种野值剔除准则在目标预测中的应用研究[J].指挥控制与仿真,2011,33(4):98-102
Lu Y L, He J Z, An J, Miao G J. Research on rules for eliminating outliers and its application to target prediction[J]. *Command Control & Simulation*, 2011, 33(4): 98-102 (in Chinese)

[16] Zhang M L, Zhou Z H. A review on multi-label learning algorithms[J]. *IEEE Transactions on Knowledge and Data Engineering*,2014,26(8):1819-1837

[17] Utgoff P E,Berkman N C,Clouse J A. Decision tree induction based on efficient tree restructuring[J]. *Machine Learning*, 1997,29(1):5-44

[18] 周志华.机器学习[M].北京:清华大学出版社,2016:73-74
Zhou Z H. *Machine Learning* [M]. Beijing: Tsinghua University Press,2016:73-74 (in Chinese)

[19] Weinberger K Q,Blitzer J,Saul L K. Distance metric learning for large margin nearest neighbor classification [J]. *Journal of Machine Learning Research*,2009,10(1):207-244

[20] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss[J]. *Machine Learning*, 1997,29(2-3):103-130

[21] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. *Machine Learning*,1997,29(2-3):131-163

[22] Grossman D, Domingos P. Learning Bayesian network classifiers by maximizing conditional likelihood [C]. In: *Proceedings of the 21th International Conference on Machine Learning* 2004. New York:Association Computing Machinery, 2004:361-368

[23] Bradley A P. The use of the area under the ROC curve in the evaluation of machine learning algorithms [J]. *Pattern Recognition*,1997,30(7):1145-1159

[24] Fawcett T. An introduction to ROC analysis [J]. *Pattern Recognition Letters*,2006,27(8):861-874

[25] Hand D J,Till R J. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. *Machine Learning*,2001,45(2):171-186

责任编辑:刘迎春