

基于深度学习的白酒分类识别方法

Research on the application methods of classifying and recognizing Baijiu wine based on deep learning

刘 鑫^{1,2} 韩 强^{1,2,3} 周永帅^{1,2} 度先国^{1,2}

LIU Xin^{1,2} HAN Qiang^{1,2,3} ZHOU Yong-shuai^{1,2} TUO Xian-guo^{1,2}

(1. 四川轻化工大学自动化与信息工程学院, 四川 自贡 643000; 2. 人工智能四川省

重点实验室, 四川 自贡 643000; 3. 泸州老窖集团有限责任公司, 四川 泸州 646000)

(1. School of Automation and Information Engineering, Sichuan University of Light Chemical Technology, Zigong, Sichuan 643000, China; 2. Sichuan Provincial Key Laboratory of Artificial Intelligence, Zigong, Sichuan 643000, China; 3. Luzhou Laojiao Group Co., Ltd., Luzhou, Sichuan 646000, China)

摘要:以深度学习为基础,结合 Tensorflow 与 Keras 框架,建立了基于深度学习的白酒品牌分类预测模型。通过电子舌(阵列式传感器)对待测白酒的特征信息进行采集,并与已知的待测白酒样品类别结合建立测试样本数据集,通过训练集与测试集对基于深度学习的白酒品牌分类预测模型进行训练与性能检验。结果表明,该预测模型的白酒品牌识别率达 99.987%,准确率较高。

关键词:深度学习;分类预测;电子舌;实用性;高效性

Abstract: A deep learning-based model for the classification and prediction of Baijiu wine brands is developed, combining with the Tensorflow and Keras frameworks. The test sample data set is established by collecting the feature information of Baijiu wine to be tested through an electronic tongue (an array sensor), combining the known Baijiu wine categories with that of the unknown one (to be test). Then the deep learning-based Baijiu wine brand classification prediction model was trained and tested for, preparing for the performance by using the trained and tested sets. The results showed that the prediction model achieves 99.987% of Baijiu wine brand recognition rate, showing a high accuracy.

Keywords: deep learning; classification prediction; electronic tongue; practicality; high efficiency

白酒的分类识别技术是白酒研发生产中较为重要的

一项技术,其性能有助于提升酒企的生产效率,如区别不同品牌的酒,同一品牌不同层次的酒以及不同等级的基酒。传统的白酒检测主要依靠品酒师的感官品评经验^[1],主观性强,不能实时、快速、准确地提供检测数据。随着人工智能技术的高速发展,神经网络、模式识别等技术逐渐被运用至白酒领域,电子鼻、电子舌等智能仿生检测设备^[2]也随之诞生。在先进检测设备和技术广泛应用的背景下,白酒的相关数据迅速扩增,导致传统的白酒分类识别方法^[2-3](如人工品评、SVM 等)无法满足分类识别的具体需求,其准确率低、耗时长等弊端逐渐显露。近年来,气相(GC)、气质联用(GC-MS)^[4]等仪器被应用至白酒的数据测试中,此类检测方法的鉴别精度极高,但因其存在操作复杂、检测时间长等缺点,仅能在实验室进行小样检测。

深度学习是指基于多层人工神经网络的机器学习方法,因其对高维、冗余的数据有极大的优势而在各个领域得到了广泛的应用。基于深度学习的分类预测应用仅是其中的一个功能领域,其中文本分类、图像分类、语音识别等的应用^[5-7]已较为成熟。文章拟使用阵列式传感器(电子舌)采集白酒特征数据结合深度学习^[8]建立白酒品牌分类识别模型,以期实现不同品牌白酒的高效、精准分类。

1 基于深度学习的白酒分类预测模型

在深度学习的分类应用中,数据集的质量与模型最终的分类预测准确率息息相关。高质量的数据集可以训练出高性能的分类预测模型,能够对待测数据类别进行精准预测。通常情况下,收集的数据集不能直接使用,需要进行预处理,包括数据筛选(异常数据处理、无效数据删除等)、数据标签(数据类别)设定、数据标准化以及训

基金项目:四川省科技计划项目(编号:2021YFS0339);四川省科技成果转化示范项目(编号:2020ZHCG0040);四川省重大科技专项项目(编号:2018GZDZX0045)

作者简介:刘鑫,男,四川轻化工大学在读硕士研究生。

通信作者:韩强(1987—),男,四川轻化工大学讲师,博士。

E-mail: hanqiang1117@163.com

收稿日期:2020-11-13

训练集和测试集的划分。其中,数据标准化^[9]的方法有:Min-max 标准化和 Z-score 标准化。其目的是对原始数据进行线性变换,使结果落在 $[0,1]$ 或 $[-1,1]$ 区间,取消由于量纲不同、自身变异或者数值相差较大所引起的误差。

Min-max 对应的表达式为:

$$x^* = \frac{x - \min}{\max - \min}, \quad (1)$$

式中:

max——同类样本数据最大值;

min——同类样本数据最小值。

Z-score 标准化对应的表达式为:

$$x^* = (x - \mu) / \sigma, \quad (2)$$

式中:

μ ——所有样本数据均值;

σ ——所有样本数据标准差。

Z-score 标准化会改变原有数据的分布结构,比较适合类正态分布的数据。因此,选择 Min-max 标准化将所测数据全部处理至 0~1。训练集与测试集的划分是在原始数据标准化的基础上进行,参照各类经典案例中的分配情况及自身的试验经验,训练集设为样本总数的 70%,剩余的归为测试集。为保障测试集与训练集内部数据分布均匀(各类别所含比例相当)且具有随机性,在数据集划分前需对总样本进行乱序调整。深度学习的模型结构是依据标准化后的样本数据特征及分类的类别来确定,数据特征对应神经网络输入层神经元的个数,分类的类别数对应输出层神经元的个数,中间层神经元的个数依据设定规则进行初步的设定并结合模型分类预测的效果进行反馈调节。各类参数确定后,进行模型训练,在神经网络的运算过程中,核心步骤是前向传播与反向传播^[10],其中,前向传播表达式为:

$$A^l = g(Z^l) = g(A^{l-1} * W^l + b^l), \quad (3)$$

式中:

l ——网络层数;

W ——权值向量;

$*$ ——卷积运算;

b ——偏置量 bias;

g ——激活函数。

反向传播的主要目的是更新权值与偏置,其表达式为:

$$\begin{aligned} W^{[l]} &= W^{[l]} - \alpha \left(\frac{1}{m} dZ^{[l]} * A^{[l-1]T} \right) \\ b^{[l]} &= b^{[l]} - \alpha \left[\frac{1}{m} n p. \text{sum}(dZ^{[l]}) \right], \end{aligned} \quad (4)$$

式中:

α ——学习率;

l ——神经网络层数。

为增加模型的非线性特性,在隐含层和输出层的输出运算中引入激活函数。常见的激活函数^[11-12]有: Sigmoid 函数、Tanh 函数、Relu 函数、Logistic 函数和 Softmax 函数。Logistic 与 Softmax 主要在分类中使用,其中 Logistic 具体针对的是二分类问题,Softmax 主要解决的是多分类问题。Sigmoid、Tanh 及 Relu 通常被用为隐含层激活函数,其最大的区别在于函数值域不同。结合试验数据特征,隐藏层选择 Relu 作为激活函数,由于数据类别超过两类,故输出层选择 Softmax 作为激活函数,其数学表达式为:

$$\sigma_i(z) = \frac{\exp(z_i)}{\sum_{j=1}^m \exp(z_j)}, \quad (5)$$

式中:

z_i ——第 i 个类别的预测结果。

模型经 Softmax 函数得出的结果为概率值,其最大值所对应的类别即为预测结果(类别)。损失函数作为模型性能评价重要方法,在深度学习的分类应用中,交叉熵损失函数(Cross Entropy)应用最为广泛。其中 Binary Cross Entropy 主要运用在二分类领域中,Softmax 结合 Cross Entropy 主要针对多分类问题,其表达式为:

$$H(p, q) = - \sum_{i=1}^n p(x_i) \log[q(x_i)], \quad (6)$$

式中:

$p(x_i)$ ——真实概率分布;

$q(x_i)$ ——预测概率分布。

由于试验数据来源于阵列式传感器(电子舌)对待测白酒的实时检测,数据量较大。为了加快深度学习分类预测模型的训练速度^[13],采用 Mini-batch 方法将训练数据均分成多份,每次学习一份数据,小批量的进行梯度下降,按批更新参数,同一批次的的数据共同决定了本次梯度的方向,减少了梯度下降过程中的随机性,与此同时,每份数据相对于整个数据集来说模型的计算量变小,加快了模型训练速度。在模型的训练过程中还引入 Adam 优化器^[14]对模型进行优化。Adam 优化器结合了 AdaGrad 和 RMSProp 两种优化算法的优点,在自动调整学习率的同时保证参数的更新不受梯度伸缩变换的影响,其参数更新的数学表达式为:

$$\theta_t = \theta_{t-1} - \alpha * \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon), \quad (7)$$

式中:

$\hat{m}_t$ ——梯度均值;

$\hat{v}_t$ ——梯度方差。

由式(7)可知,对更新的步长进行计算,能够从梯度均值及梯度平方两个角度进行自适应调节,而不是直接由当前梯度决定,能够达到提升模型性能的目的。在深度学习分类预测模型的训练过程中,极易出现过拟合现象,因此采用 Dropout 方法(参数正则化)解决过拟合问

题^[15-16],Dropout 可作为训练深度神经网络的一种 trick 供选择,在每个训练批次中,通过忽略一定数量的特征检测器(使一定数量的隐层节点值为 0),可以明显地减少过拟合现象。同时,神经元数量的减少进一步加快了深度学习模型的训练速度。基于深度学习分类应用的流程图如图 1 所示。

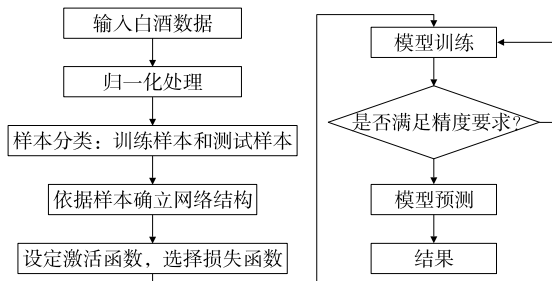


图 1 深度学习分类应用流程图

Figure 1 Flow chart of deep learning classification application

2 实验验证

2.1 试验装置

采用多频脉冲阵列式传感器(电子舌系统),其中传感器阵列由铂、银、钨、钛、钡、金圆盘工作电极构成,直径 Φ 均为 2 mm,工作温度 15~35℃,工作湿度 0%~80%。设备开机后需在 0.01 mol/L 氯化钾溶液中对传感器预热 30 min,分别设置系统的起始电压、结束电压和步降降压值为 1.0,-1.0,0.2 V,其中传感器提取数据的脉冲频率设为 100 Hz。

2.2 数据获取

选用电子舌的人工进样系统数据采集模型对不同品牌的白酒进行特征信息采集。进样机共有 12 个柱形槽口用于放置盛有待测液体的容器,可以通过软件控制传感器测试待测液体位置,其中 1 号位必须为清洗液且清洗液可放置多个。阵列传感器 S1~S6 的灵敏度可依据样品的不同进行调整,灵敏度的选择一般使信号最大值在 0.5~10.0 V 为宜,信号值过大需调小灵敏度,信号值过小则需适量调大灵敏度。信号过大或过小都可能造成不能保存的数据的失误。样品经一次采集后,会产生 1×6 维的数据,试验选取市面上常见的多个品牌进行测试,其中部分品牌的酒样见表 1。为了保证样品采集的随机性,分别对每类酒进行若干次采集而非固定采集次数,初步得到每个品牌平均约 10 000 组数据,最后构成样本数据集,通过数据库进行存储。从全部数据中选取 8 类 6 维数据作为测试样本,8 类酒分别用数字 1,2,3,4,5,6,7,8 进行标识分类,每一类数据具有 6 个特征值。所选数据的部分值见表 2,传感器采集的所有数据的分布图如图 2 所示。

表 1 部分待测品牌样品酒

Table 1 Some sample wines of different brands to be tested

标识类别	品牌	产地	酒精度/%Vol	数量/瓶
1	伊力特	新疆昌黎	46	4
2	稻花香	湖北宜昌	52	4
3	红星二锅头	北京	56	4
4	泸州老窖	四川泸州	52	4
5	五粮液	四川宜宾	52	4
6	舍得	四川射洪	52	4
7	江津老白干	重庆	50	4
8	西凤酒	陕西	52	4

表 2 部分样品值

Table 2 Some sample values

类别	S1	S2	S3	S4	S5	S6
1	1.396 8	-2.030 0	-1.060 0	5.450 0	1.057 8	-1.250 0
1	1.417 8	-2.040 0	-1.070 0	5.380 0	1.062 8	-1.290 0
1	1.408 0	-2.050 0	-1.080 0	5.270 0	1.064 7	-1.300 0
2	1.290 0	-1.870 0	-9.990 0	5.030 0	9.856 1	-1.290 0
2	1.875 6	-2.090 0	-1.460 0	6.430 0	1.429 1	-6.110 0
2	1.880 0	-2.020 0	-1.490 0	6.200 0	1.431 0	-6.770 0
3	1.850 2	-2.000 0	-1.480 0	6.090 0	1.426 6	-6.850 0
3	1.800 2	-1.700 0	-1.430 0	5.780 0	1.388 1	-3.350 0
3	1.049 0	-1.820 0	-7.280 0	4.610 0	7.570 7	-3.540 0
4	1.055 0	-1.600 0	-7.800 0	4.290 0	7.842 2	-3.390 0
4	1.069 2	-1.650 0	-7.950 0	4.250 0	7.953 6	-3.540 0
4	1.524 8	-1.330 0	-6.010 0	2.120 0	3.099 8	-7.950 0
5	1.253 7	-3.125 0	-2.371 0	3.542 0	0.623 1	-2.345 0
5	1.358 4	-3.584 0	-2.887 0	3.965 7	0.984 1	-2.456 0
5	1.412 5	-3.728 6	-2.786 7	3.796 3	0.886 5	-2.678 0
6	2.132 1	-2.493 4	-1.891 0	7.435 0	1.100 2	-5.632 1
6	2.034 7	-2.400 0	-1.803 0	7.987 0	1.103 4	-5.354 7
6	1.983 4	-2.217 0	-1.834 0	7.035 2	1.003 9	-5.038 1
7	0.741 6	-3.234 0	-2.362 0	5.862 0	6.734 1	-5.193 1
7	0.740 1	-3.321 3	-2.300 1	5.698 7	6.637 4	-10.416 7
7	0.693 4	-3.104 5	-2.134 0	5.563 4	6.874 7	-5.184 7
8	1.792 1	-2.389 4	-3.216 3	3.793 2	2.471 3	-9.834 7
8	1.754 3	-2.376 1	-3.200 4	3.805 1	2.208 4	-9.620 4
8	1.702 1	-2.205 4	-3.178 9	3.682 1	2.108 9	-9.074 8

2.3 数据处理与模型参数设定

在阵列式传感器(电子舌)所测的所有数据中选取 8 类数据进行异常值、无效数据删除等预处理操作,最终确立类别均匀的 50 000 组数据作为深度学习分类预测模型的数据集。由于模型的损失函数采用的是交叉熵,需对数据集在[0,1]区间内进行归一化处理,再按照 30%,70%的比例对数据集进行测试集与训练集的划分,划分数据集需注意各种类别在训练集与测试集中分布均匀。

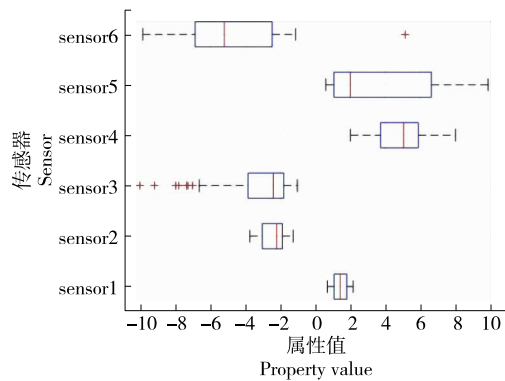


图2 样品在各个属性上的分布

Figure 2 Distribution of samples on various attributes

载入模型中进行训练前还需作乱序处理,以保证数据的随机性。依据阵列式传感器(电子舌)所测数据的维度、白酒品牌的类别与隐含层的设计经验,结合初步结果进行反馈调整,模型的结构最终设定为6—64—8。为避免模型过拟合,Dropout的值设为0.5,即忽略1/2的特征检测器。模型的迭代次数并不是越大越好,结合模型结果,经过实际训练调整,最终设定该模型最多迭代次数为70。

2.4 结果分析

在深度学习分类预测模型中,经过训练集对所建模型进行训练,通过测试集进行预测,得到预测样本识别率为99.987%。其损失函数效果见图3,正确率见图4。深

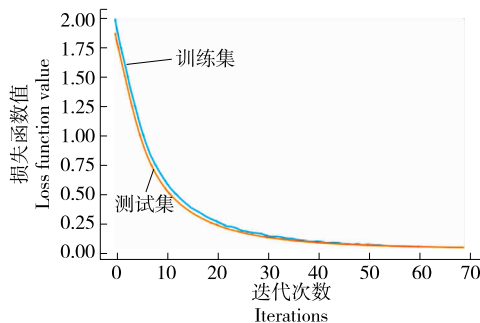


图3 训练模型损失函数效果图

Figure 3 Effect diagram of training model loss function

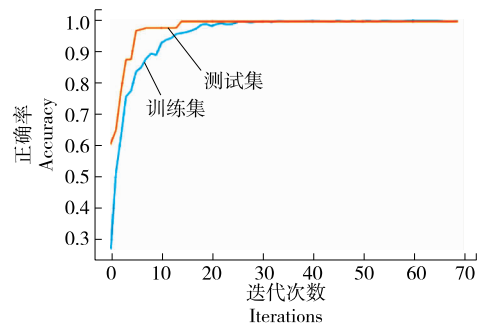


图4 测试样本训练集与测试集的正确率

Figure 4 Correct rate of test sample training set and test set

度学习分类预测模型经训练后,交叉熵损失函数的值从开始训练的2.0704降低至训练结束的0.0879,说明模型的性能良好,能够满足分类需求。

由图4可知,训练集与测试集的正确率均较高,且未出现过拟合现象,能够满足预测的精度要求,达到精准分类的目的。

基于深度学习的分类识别与传统不同核函数的SVM分类识别方法对比结果如表3所示。由表3可知,传统的分类识别方法(不同核函数的SVM)对于白酒品牌分类准确率较低,而使用深度学习建立的白酒分类识别方法准确率较高,实用性更好。

表3 不同方法识别率比较

Table 3 Comparison of recognition rate of different methods

方法	核函数	分类准确率/%
SVM	线性核	78.260
SVM	多项式核	85.860
SVM	RBF	88.040
深度学习		99.987

3 结论

文章提出的基于深度学习的白酒分类预测模型将深度学习引入至白酒领域中,再结合阵列式传感器(电子舌)对白酒快速、高效的检测优势,能够高效、精准地对不同品牌的白酒进行分类识别。试验结果表明,该预测模型的白酒品牌识别率达99.987%,具有较强的非线性特性,能够充分发掘数据特征与白酒类别间的关系,具有较好的实用性。后续将提升基于深度学习的白酒分类预测模型的速度,并将深度学习等更多的人工智能技术引入至白酒领域中,实现白酒的智能酿造、生产、勾兑以及生产状态实时监测等目标。

参考文献

- [1] 蒋鼎国,周红标,耿忠华.基于PSO-SVM的白酒品质鉴别电子鼻[J].中国酿造,2011(11):149-152.
- [2] 谢丽慧.基于SVM算法的智鼻与智舌技术在食用植物油综合质量检测中的应用[D].杭州:浙江工商大学,2015:34-53.
- [3] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: A local SVM approach [C]// International Conference on Pattern Recognition. Cambridge, United Kingdom: IEEE, 2004: 32-36.
- [4] 罗珠,黄箭,李杨华,等. GC-MS在白酒各类物质分析中的研究进展[J].酿酒科技,2018(12):117-119.
- [5] 杜思佳,于海宁,张宏莉.基于深度学习的文本分类研究进展[J].网络与信息安全学报,2020,6(4):1-13.

(下转第79页)

物油中游离棉酚含量测定的扩展不确定度 $U(x) = 2.3 \times 2 = 4.6 \text{ mg/kg}$ 。

根据 GB 5009.148—2014 中检出结果有效保留位数等要求,最终植物油中游离棉酚含量测定结果的报告为:

$$C = (48 \pm 5) \text{ mg/kg } (K=2)。$$

5 结论

通过对植物油中游离棉酚含量的检测,分析和评定了游离棉酚检测过程中的不确定度主要来源,并计算出检测结果的扩展不确定度。结果表明,植物油中游离棉酚含量检测结果不确定度的重要来源为高效液相色谱仪的性能、样品前处理的重复性以及标准曲线拟合。因此在植物油中游离棉酚检测过程中,要做好高效液相色谱仪的性能维护与监控并加以控制;完善标准曲线的配置,做好标准物质的有效核查;加大检测人员培训,完善和规范检测程序,确保人员操作的规范性和一致性。

参考文献

[1] CORINNA K, HANNA M, HEIDEL F, et al. Gossypol toxicity and detoxification in *Helicoverpa armigera* and *Heliothis virescens*[J]. *Insect Biochemistry and Molecular Biology*, 2016, 78: 69-77.

[2] 中华人民共和国国家卫生健康委员会,国家市场监督管理总局. 食品安全国家标准 植物油: GB 2716—2018[S]. 北京: 中国标准出版社, 2018: 1-3.

[3] 中国标准化委员会. 化学分析测量不确定度评定: JJF 1135—2005[S]. 北京: 中国标准出版社, 2005: 1-12.

[4] 中国国家标准化管理委员会. 测量不确定度的评定与表示通

用要求: GB/T 24718—2017[S]. 北京: 中国标准出版社, 2017: 1-76.

[5] 唐吉旺, 何浩, 王淑霞, 等. 固相萃取—高效液相色谱—串联质谱法同时测定动物源性食品中五氯酚和游离棉酚[J]. *食品科技*, 2020, 45(7): 364-371.

[6] 王雅朦, 郭咪咪, 魏征, 等. 食用植物油中痕量游离棉酚的超高效液相色谱—串联质谱测定[J]. *中国粮油学报*, 2019, 34(4): 126-132.

[7] 邸万山. 高效液相色谱内标法测定棉籽油中游离棉酚[J]. *中国油脂*, 2016(6): 97-99.

[8] 中国认证认可监督管理委员会. 检验检测机构资质认定能力评价 检验检测机构通用要求: RB/T 214—2017[S]. 北京: 中国标准出版社, 2017: 1-10.

[9] 中国合格评定国家委员会. 测量不确定度的要求: CNAS-CL01-G003—2019[S]. 北京: 中国标准出版社, 2019: 1-6.

[10] 中国标准化委员会. 测量不确定度评定与表示: JJF 1059.1—2012[S]. 北京: 中国标准出版社, 2012: 1-53.

[11] 中国合格评定国家委员会. 测量不确定度要求的实施指南: CNAS-GL05—2011[S]. 北京: 中国标准出版社, 2011: 1-7.

[12] 中国合格评定国家委员会. 化学分析中不确定度的评估指南: CNAS-GL06—2019[S]. 北京: 中国标准出版社, 2019: 1-137.

[13] 袁河, 肖晓义, 刘佳, 等. 高效液相色谱法测定食用槟榔中苯甲酸、山梨酸和糖精钠的不确定度评定[J]. *食品与机械*, 2020, 36(8): 77-80.

[14] 黄坤, 王会霞, 范小龙, 等. 超高效液相色谱—串联质谱法测定炒货中组胺含量的不确定度评定[J]. *食品与机械*, 2019, 35(8): 64-68.

(上接第 71 页)

[6] 苏赋, 吕沁, 罗仁泽. 基于深度学习的图像分类研究综述[J]. *电信科学*, 2019, 35(11): 58-74.

[7] 侯一民, 周慧琼, 王政一. 深度学习在语音识别中的研究进展综述[J]. *计算机应用研究*, 2017, 34(8): 2 241-2 246.

[8] 赵为. 基于深度学习的电子鼻白酒识别方法研究[D]. 天津: 天津大学, 2018: 29-51.

[9] 邱锡鹏. 神经网络与深度学习[J]. *中文信息学报*, 2020, 34(7): 4.

[10] 张宪超. 深度学习: 上[M]. 北京: 科学出版社, 2019: 1-37.

[11] 张向荣. 模式识别[M]. 西安: 西安电子科技大学出版社, 2019: 89-103.

[12] 万磊, 佟鑫, 盛明伟, 等. Softmax 分类器深度学习图像分类方法应用综述[J]. *导航与控制*, 2019, 18(6): 1-9, 47.

[13] 高晗, 田育龙, 许封元, 等. 深度学习模型压缩与加速综述[J]. *软件学报*, 2021, 32(1): 68-92.

[14] BYUNG J K. Improved deep learning algorithm[J]. *Journal of Advanced Information Technology and Convergence*, 2018, 8(2): 119-127.

[15] 李恒. 基于深度学习过拟合现象的分析[J]. *中国科技信息*, 2020(14): 90-91.

[16] 韩梦娇. 基于选择性区域丢弃的 dropout 方法研究与实现[D]. 成都: 电子科技大学, 2020: 21-36.

(上接第 75 页)

[11] D21 Initiative. D21-qualitätskriterien für internetangebote[EB/OL]. (2018-04-20) [2021-04-13]. https://initiated21.de/app/uploads/2019/08/d21-qualitaetskriterien_2018.pdf.

[12] 中华人民共和国商务部网站. 德国网络零售情况[EB/OL]. (2012-05-16) [2021-04-13]. <http://www.mofcom.gov.cn/aarticle/i/dxfw/jlyd/201205/20120508127788.html>.

[13] 罗辉. 中国食品安全监管的问题审视与机制向度[J]. *食品与机械*, 2019, 35(8): 100-103.

[14] 李文娟, 刘桂锋, 卢章平. 基于专利分析的我国大数据产业技术竞争态势研究[J]. *情报杂志*, 2015(7): 65-70.

[15] HOOKER C N H. HACCP as an international trade standard[J]. *American Journal of Agricultural Economics*, 1996, 78(3): 775-779.