宽度学习系统在蘑菇毒性判别中的应用

李旺，俞祝良

（华南理工大学自动化科学与工程学院，广东广州 510640）

摘要：为了提升蘑菇毒性判别的准确率，消除个体差异，本文提出了一种基于宽度学习系统的蘑菇毒性判别方法。文章首先对蘑菇各特征指标与其毒性判别的相关性进行了探究，其所得结果显示蘑菇的气味和颜色是其区分度最大的特征，该结果与人工判别积累的经验相符。接着构建宽度学习系统并进行训练，对比分析不同样本量情况下所提方法的准确率，发现当样本数量大于 1000 时，宽度学习系统分类准确率便高于 99.5%。与 BP-神经网络相比，该方法准确率高且所需训练时间短。最后依据宽度学习系统的增量学习算法，在模型性能不满足要求时，通过增加隐藏层节点快速更新系统，使系统分类准确率从 98.55%提升至 99.99%，而不需要将网络重新进行训练，这使实时判别蘑菇毒性成为可能。因此对比其他方法，基于宽度学习系统的蘑菇毒性判别方法具有准确率高、训练时间短、判别迅速且易拓展的优点。

关键词：毒蘑菇；机器学习；宽度学习系统

文章篇号：1673-9078(2019)07-267-272 DOI: 10.13982/j.mfst.1673-9078.2019.7.037

Application of Broad Learning System in Discrimination of Mushroom Toxicity

LI Wang, YU Zhu-liang

(College of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China)

Abstract: In order to improve the accuracy of mushroom toxicity discrimination and eliminate individual differences, a method of mushroom toxicity discrimination based on broad learning system was proposed in this work. Firstly, the correlation between each characteristic of mushroom and its toxicity was explored. The results showed that the odor and color of mushroom were the most distinguishing characteristics. These results were consistent with the experience of manual discrimination. Then, broad learning system was established and trained. By comparing performance in diverse sample sizes of different methods, it was found that when the sample size is larger than 1000, the classification accuracy of the broad learning system was higher than 99.5%. Compared with BP-neural network, the proposed method was of high-accuracy and fast-training. Finally, according to the incremental learning algorithm of broad learning system, when the performance of the system does not meet the requirements, the system can be updated quickly by increasing the hidden nodes, and the accuracy can be improved from 98.55% to 99.99% without retraining the whole network. It was possible to discriminate the toxicity of mushrooms in real time. Therefore, compared with other methods, this method of mushroom toxicity discrimination based on broad learning system has the advantages of high accuracy, short training time, rapid discrimination and easy expansion.

Key words: poisonous mushroom; machine learning; broad learning system

蘑菇（*Agaricus campestris*），一种常见的菌类植物，广泛分布于地球各处，在森林落叶地带最为丰富。蘑菇营养丰富，富含人体内必须的氨基酸、维生素等营养成分，是一种广受欢迎的营养保健食品。但有不少蘑菇生长在阴暗、潮湿地带，其在生长过程中吸收了毒素，不再适合人食用。因毒蘑菇与蘑菇外形相似，难以分辨，我国每年都会有误食毒蘑菇事件发生，特别是春夏两季，严重时甚至致人死亡。2001 年 9 月 1 日发生在江西永修县的毒蘑菇中毒事件是新中国成立以来最大的毒蘑菇中毒事件，上千人因毒蘑菇中毒。因此，如何根据蘑菇的各种特性来判别蘑菇是否有毒的方法是非常重要的[1,2]，有助于减少蘑菇中毒事件的发生，进一步保障食品安全。

民间判别蘑菇是否有毒一般依据蘑菇的生长地带、颜色、外形、分泌物和气味等特征来进行判断，但这种判别及其依赖个人经验，且大部分经验只适用

收稿日期：2019-04-22

基金项目：广东省科技计划项目（2018B010107001）

作者简介：李旺（1993-），女，硕士研究生，研究方向：机器学习与模式识别

通讯作者：俞祝良（1973-），男，博士，教授，研究方向：模式识别、机器学习、信号处理等

于部分地区蘑菇，判别准确性有待商榷，广泛推广有一定的限制[1]。而在学术界，通常通过研究蘑菇的毒性成分、中毒机理来判别蘑菇毒性[2]，此种方法虽然准确率高，但因检测成本高、实验条件要求多等缺点，难以推广到工程应用。

近几年来，机器学习作为人工智能的核心，在图像识别、语音识别领域迅猛发展，它作为一门多领域交叉学科，在食品安全领域的应用也越来越受到科学家的关注[3,4]。蘑菇毒性的判别问题本身就是一种典型的多维度、非线性分类问题。将机器学习方法应用于蘑菇毒性判别是一个值得研究的方向。

宽度学习系统(Broad Learning System)[5]是澳门大学陈俊龙教授于 2018 年初在传统的随机向量链接神经网络(Random Vector Functional-Link Neural Network，RVFLNN)[6]的基础上提出的，相比于卷积神经网络等深度神经网络，其在解决非线性、中小样本的问题上有独特的优势。

本文提出了一种基于宽度学习系统的蘑菇毒性判别方法，旨在建立一个准确、简洁、可靠的能应用于生产线上蘑菇毒性自动化检测和判别系统。

1 基于宽度学习系统的蘑菇毒性判别方法

1.1 宽度学习系统简介

宽度学习系统是在传统的 RVFLNN 的基础上提出的，假设有 K 个 D 维的输入数据 $\boldsymbol{X}\in\boldsymbol{R}^{K\times D}$，输入数据通过式（1）的映射可得到 n 组映射特征，每组包含节点 v 个。

$$\boldsymbol{Z}_i=\phi(\boldsymbol{X}\boldsymbol{W}_{ei}+\boldsymbol{\beta}_{ei}),\ i=1,2,\ldots,n \tag{1}$$

其中：$\boldsymbol{W}_{ei}$ 和 $\boldsymbol{\beta}_{ei}$ 为随机生成的权重，令 $\boldsymbol{Z}^n=[\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_n]$ 为映射形成的前 n 组映射特征集合，特征节点可经过激活函数非线性变换式（2）生成 m 组增强节点，每组包含节点 η 个。

$$\boldsymbol{E}_j=\zeta(\boldsymbol{Z}^n\boldsymbol{W}_{h_j}+\boldsymbol{\beta}_{h_j}),\ j=1,2...,m \tag{2}$$

其中：$\boldsymbol{W}_{h_j}$ 和 $\boldsymbol{\beta}_{h_j}$ 也是随机生成的权重。

因此，宽度学习系统由式（3）表示为

$$\begin{aligned}\boldsymbol{Y}&=[\boldsymbol{Z}_1,...,\boldsymbol{Z}_n|\zeta(\boldsymbol{Z}^n\boldsymbol{W}_{h_1}+\boldsymbol{\beta}_{h_1}),...\zeta(\boldsymbol{Z}^n\boldsymbol{W}_{h_m}+\boldsymbol{\beta}_{h_m})]\boldsymbol{W}\\&=[\boldsymbol{Z}_1,...,\boldsymbol{Z}_n|\boldsymbol{E}_1,...,\boldsymbol{E}_m]\boldsymbol{W}\\&=[\boldsymbol{Z}^n|\boldsymbol{E}^m]\boldsymbol{W}\\&=\boldsymbol{H}\boldsymbol{W}\end{aligned} \tag{3}$$

$\boldsymbol{Y}\in\boldsymbol{R}^{K\times Q}$ 指网络的 Q 维输出。当只考虑二分类任务时，$\boldsymbol{Y}\in\boldsymbol{R}^{K}$。宽度学习系统的结构示意图如下所示。

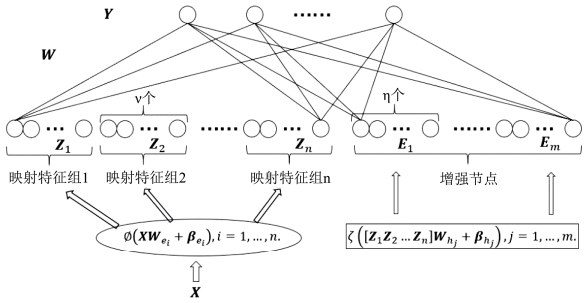


图 1 宽度学习网络的结构示意图

Fig.1 The standard structure of broad learning system

1.2 宽度学习系统求解方法

不同于其他单隐层神经网络应用梯度下降法调整各参数，在宽度学习网络中，其输入权重 $\boldsymbol{W}_e$ 和 $\boldsymbol{W}_h$ 是随机选择的，当输入权重确定后，则其不再是一个变量，无需通过训练求解或调参。因此宽度学习系统只需求解输出权重 $\boldsymbol{W}$。其求解问题可由公式（4）表示：

$$\min_{\boldsymbol{W}} f(\boldsymbol{W})=\min_{\boldsymbol{W}}\|\boldsymbol{H}\boldsymbol{W}-\boldsymbol{Y}\|_F^2 \tag{4}$$

其中：$\|\cdot\|_F$ 是指 l_F 范数。$\boldsymbol{H}$ 是宽度学习系统的隐藏层：

$$\boldsymbol{H}=[\boldsymbol{Z}^n|\boldsymbol{E}^m]=(\boldsymbol{h}_1,\boldsymbol{h}_2,\ldots,\boldsymbol{h}_L)\in\boldsymbol{R}^{K\times L} \tag{5}$$

式（4）是一个最小二乘问题，是关于 $\boldsymbol{W}$ 的凸优化估计，旨在求出最小训练误差时的输出权重。

对式（4）进行求解，可以得到：

$$\boldsymbol{W}=\boldsymbol{H}^{+}\boldsymbol{Y} \tag{6}$$

其中：$\boldsymbol{H}^{+}$是 $\boldsymbol{H}$ 的伪逆。$\boldsymbol{h}_j\in\boldsymbol{R}^{K}$，$j=1,\ldots,L$ 为隐藏层 $\boldsymbol{H}$ 的第 j 节点，K 为输入样本数量，$L=nv+m\eta$是隐藏层节点数。

但通常情况下，上述解的泛化误差可能会很大，特别是对于一些病态问题。因此为了提升系统的泛化能力，我们在原式（4）上加上 l_2 范数正则项防止网络过拟合：

$$\min_{\boldsymbol{W}} f(\boldsymbol{W})=\min_{\boldsymbol{W}}\|\boldsymbol{H}\boldsymbol{W}-\boldsymbol{Y}\|_F^2+C\|\boldsymbol{W}\|_F^2 \tag{7}$$

式（7）为一个经典的凸优化问题，常被成为岭回归问题。值 C 表示对 $\boldsymbol{W}$ 平方权重和的进一步约束。其闭合解被称为 Moore-Penrose 伪逆[7]。

$$\boldsymbol{W}=\begin{cases}\boldsymbol{H}^{\mathrm{T}}(C\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{-1}\boldsymbol{Y},\ K<L\\(C\boldsymbol{I}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{Y},\ K\geq L\end{cases} \tag{8}$$

因此，容易得到：

$$\boldsymbol{H}^{+}=\begin{cases}\boldsymbol{H}^{\mathrm{T}}(C\boldsymbol{I}+\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}})^{-1},\ K<L\\(C\boldsymbol{I}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}},\ K\geq L\end{cases} \tag{9}$$

1.3 宽度学习系统的增量学习算法

在某些情况下，如果网络的性能不能达到所需的

精度，宽度学习系统可以通过插入额外的增强节点来获得更好的性能[5,8]。假设现有隐藏层 $\boldsymbol{H}^m=[\boldsymbol{Z}^n|\boldsymbol{E}^m]$，当网络新增加 p 个增强节点时（如图 2 所示），相当于矩阵 $\boldsymbol{H}^m$ 增加 p 列。则相应的增加后的隐藏层为：

$$\boldsymbol{H}^{m+1}\triangleq[\boldsymbol{H}^m|\zeta(\boldsymbol{Z}^n\boldsymbol{W}_{h_{m+1}}+\boldsymbol{\beta}_{h_{m+1}})] \tag{10}$$

其中：$\boldsymbol{W}_{h_{m+1}}\in\boldsymbol{R}^{nv\times p}$，$\boldsymbol{\beta}_{h_{m+1}}\in R^p$ 分别是增加的 p 个增强节点对应的随机连接权重和偏置。

根据宽度学习系统的增强学习算法，新的隐藏层矩阵的伪逆为：

$$(\boldsymbol{H}^{m+1})^+=\begin{bmatrix}(\boldsymbol{H}^m)^+-\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}}\\ \boldsymbol{B}^{\mathrm{T}}\end{bmatrix} \tag{11}$$

其中：$\boldsymbol{D}=(\boldsymbol{H}^m)^+\zeta(\boldsymbol{Z}^n\boldsymbol{W}_{h_{m+1}}+\boldsymbol{\beta}_{h_{m+1}})$；

$$\boldsymbol{B}^{\mathrm{T}}=\begin{cases}(\boldsymbol{C})^+，\text{如果}\boldsymbol{C}\neq 0\\ (1+\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{H}^m)^+，\text{如果}\boldsymbol{C}=0\end{cases} \tag{12}$$

以及 $C=\zeta(\boldsymbol{Z}^n\boldsymbol{W}_{h_{m+1}}+\boldsymbol{\beta}_{h_{m+1}})-\boldsymbol{H}^m\boldsymbol{D}$。

于是，新的输出权重 $\boldsymbol{W}^{m+1}$ 为：

$$\boldsymbol{W}^{m+1}=\begin{bmatrix}\boldsymbol{W}^m-\boldsymbol{D}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\\ \boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\end{bmatrix} \tag{13}$$

由式(13)可以看出，该算法在增加增强节点时，仅通过计算相应节点的伪逆简单计算，就可以容易的求出新的输出权重，而不需要计算整个 $\boldsymbol{W}^{m+1}$ 的伪逆，从而可以实现快速的增量学习。

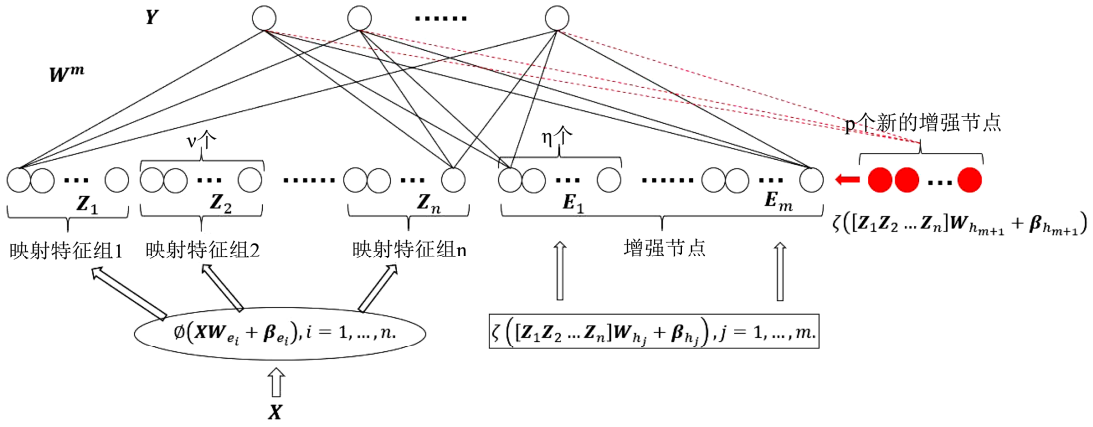


图 2 增量学习算法示意图

Fig.2 Illustration of incremental learning algorithm

表 1 部分样本的部分特征数据

Table 1 Partial characteristic data of some samples

样本编号	是否有毒	帽形	帽表面	帽颜色	是否擦伤	气味	腮-附着	腮-间距
1	e	x	y	g	t	n	f	c
2	p	x	f	g	f	f	f	c
3	p	x	f	w	f	c	f	w
4	e	f	y	n	t	n	f	c
5	e	f	f	e	t	n	f	c
6	p	x	f	w	f	c	f	w
7	p	x	s	g	f	c	f	w
8	e	x	y	e	t	n	f	c
9	p	x	f	g	f	f	f	c
10	p	x	f	g	f	c	f	c

2 实验设计及结果

实验分为两部分，均在配置 i7-8550U 2.0 GHz 的 CPU 和内存 16.0 GB 计算机的 MATLAB 环境下进行。

（1）探究蘑菇的 22 个特征指标与蘑菇毒性的相关性。

（2）使用宽度学习系统进行仿真，包括数据预处理、数据训练与测试、实验结果分析三部分。

2.1 数据集介绍

在本实验中，采用加州大学欧文分校提供的蘑菇数据集[9]进行实验，此数据集包含关于蘑菇的22个特

征，分别是：帽形、帽表面、帽颜色、是否擦伤、气味、腮-附着、腮-间距、腮-大小、腮-颜色、茎-形状、茎-根、茎-表面-上环、茎-表面-下环、茎-颜色-上环、茎-颜色-下环、菌幕-类型、菌幕-颜色、环-数量、环-类型、孢子-印记颜色、种群类别、生长地。蘑菇的这些相关特征都可以通过对其观察得到。

在本数据集中，一共有 8124 个样本，来源于奥杜邦自然保育协会推出过的关于蘑菇的指南。在所有的实验过程中，均对数据进行了数值化处理，其中用指标{-1,1}表示蘑菇是否有毒，其余各项特征类似的分别用 1,2…表示。其数据集的部分样本如表 1 所示，数值化处理后的部分样本特征如表 2 所示。

表 2 数值化后的部分样本的部分特征数据

Table 2 Partial characteristic data of some samples after digitization

样本编号	是否有毒	帽形	帽表面	帽颜色	是否擦伤	气味	腮-附着	腮-间距
1	-1	3	3	4	1	7	3	1
2	1	3	1	4	2	5	3	1
3	1	3	1	9	2	3	3	2
4	-1	4	3	1	1	7	3	1
5	-1	4	1	8	1	7	3	1
6	1	3	1	9	2	3	3	2
7	1	3	4	4	2	3	3	2
8	-1	3	3	8	1	7	3	1
9	1	3	1	4	2	5	3	1
10	1	3	1	4	2	3	3	1

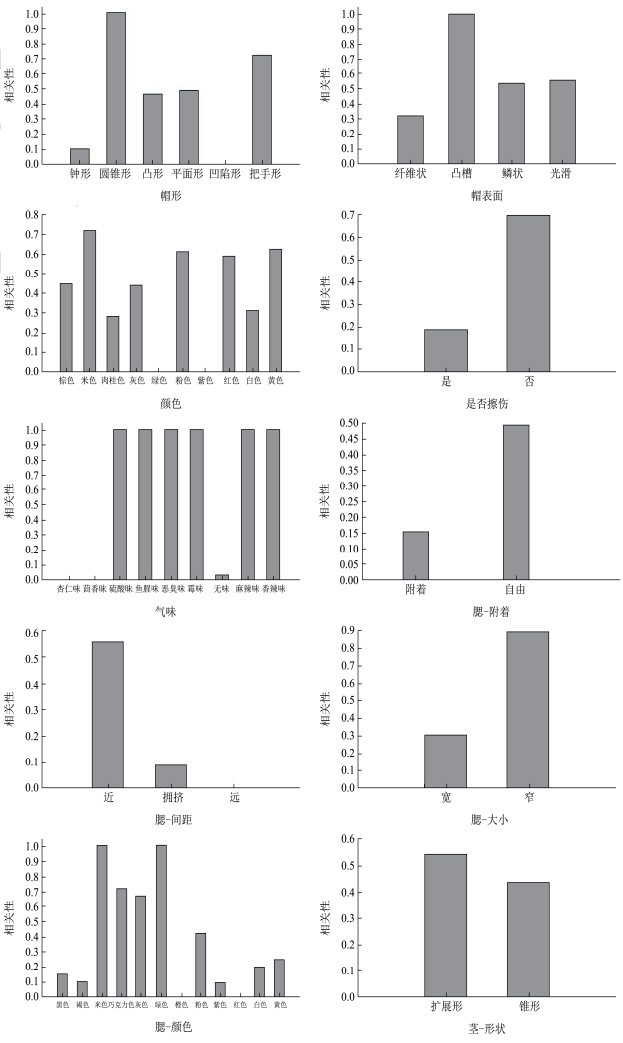
以表 1 和表 2 为例，简要说明蘑菇部分特征指标。是否有毒一列，e=edible 表示蘑菇无毒，p=poisonous 表示蘑菇有毒；帽形一列，x=convex 表示蘑菇帽形为凸形，f=flat 表是蘑菇帽形为平面形；帽表面一列，y=grooves 表示鳞状，f=fibrous 表示纤维状，s=smooth 表示表面光滑；帽颜色一列 g=gray 表示灰色，w=white 表示白色，n=brown 表示棕色，e=red 表示红色；是否擦伤一列，t=true 表有擦伤，f=false 表无擦伤；气味一列，n=none 表示无味道，f=oul 表示霉味，c=creosote 表示硫酸味；腮-间距一列，c=close 表示间距近，w=crowded 表示拥挤的。

2.2 蘑菇各指标与毒性的相关性探究

在日常生活中，判别毒蘑菇一般依靠人的经验认识，例如颜色鲜艳、形状怪异、有辛辣、酸涩等味，那么这些因素是否确实对蘑菇毒性有一定的影响呢，而这些指标与毒性之间的关系如何，在本实验中通过计算各不同指标下有毒的概率来判断单独指标与毒性之间的关系，找到能使蘑菇毒性区分度最大的特征。

22 个特征与毒性之间的相关性如下图 3 所示。

由图 3 柱状图可以看出，蘑菇的帽部特征（形状、表面特性）、气味、各部位(腮、茎、环、菌幕、孢子）颜色是其区分度最大的特征。由此可以推断出，由经验得出的根据蘑菇外形、气味、颜色等区别有毒蘑菇和无毒蘑菇的方法是合理的。



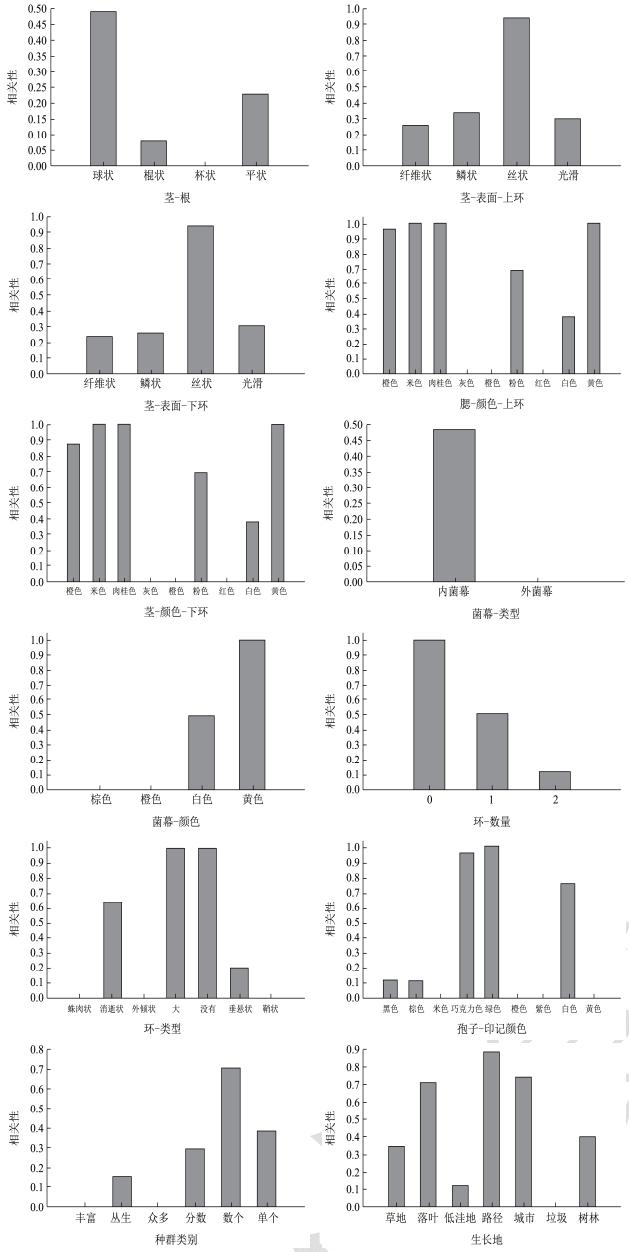


图 3 22 个特征与毒性之间的相关性

Fig.3 Correlation between 22 characteristics and toxicity

2.3 宽度学习系统在蘑菇毒性判别中的应用

使用宽度学习系统对蘑菇毒性判别进行仿真，其实验过程如图 4 所示。

在宽度学习系统中，有两个需要确定的参数是计算式（8）时的正则项参数 C 和隐藏层节点数 L，受文献[10]的启发，正则项参数 C 从集合{2^{-10}，2^{-9}，…，2^{9}，2^{10}}中选取，且采用交叉验证法[11,12]确定其最优参数 C，在保证隐藏层节点数 L 不变的情况下，参数 C 的寻优过程如图 4 所示。通过图 5 可以看到，在一定可接受范围内，正则项参数 C 对基于宽度学习系统的蘑菇毒性判别准确率的影响不大。

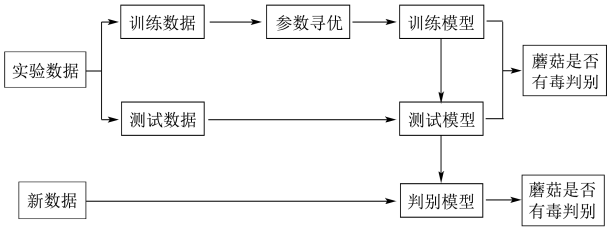


图 4 蘑菇毒性判别实验过程

Fig.4 The procedure of mushroom toxicity discrimination

在对数据进行预处理后，每次试验从样本集中随机抽取 2/3 样本作为训练集，剩余 1/3 样本作为测试集，且重复进行 20 次试验，以 20 次试验获得的准确率的平均值作为最终准确率。表 3 展示了当样本集的大小增加时，宽度学习系统和 BP-神经网络对于蘑菇毒性判别的准确率变化情况。

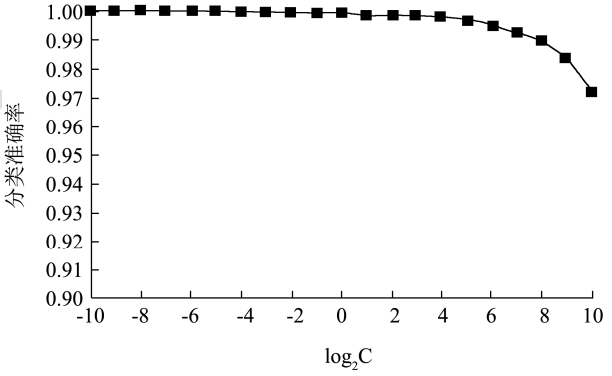


图 5 正则项参数 C 对准确率的影响

Fig.5 The influence of regularization parameter C on accuracy

表 3 各样本数量下的不同算法判别准确率

Table 3 The accuracy of different algorithms with different sample size

样本数量	宽度学习系统		BP-神经网络	
	准确率/%	训练时间/s	准确率/%	训练时间/s
500 个样本	96.89	0.0391	94.16	2.1962
1000 个样本	99.63	0.0436	99.29	2.6846
2000 个样本	99.84	0.0541	99.77	3.6257
4062 个样本	99.98	0.0737	99.97	4.8984
6000 个样本	99.99	0.0877	99.99	6.0593
8124 个样本	100	0.1112	100	7.7858

表 4 增加增强节点过程中相应的准确率

Table 4 The accuracy with different numbers of enhancement nodes

特征节点数	增强节点数	测试准确率/%
100	50	98.55
100	70	99.17
100	90	99.26
100	110	99.39
100	130	99.44
100	150	99.48
100	170	99.70
100	190	99.85
100	210	99.85
100	230	99.88
100	250	99.93
100	270	99.98
100	290	99.99
100	310	99.99
100	330	99.99

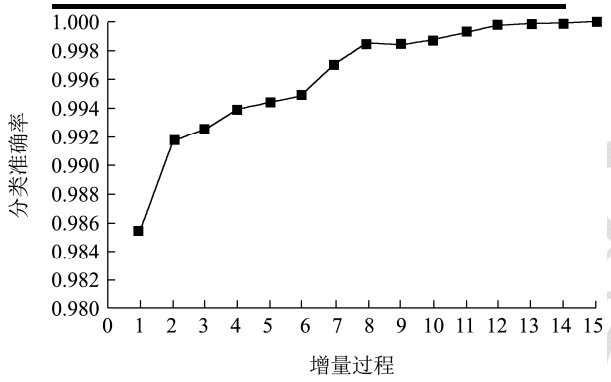


图 6 增加增强节点过程中相应的准确率折线图

Fig.6 Line chart of accuracy with different numbers of enhancement nodes

由表 3 可以看到，随着样本数量的增加，两种算法的蘑菇毒性判别的准确率随之相应增加，且宽度学习系统的判别准确率较 BP-神经网络稍高。例如，在 500 个样本这组，BP-神经网络的判别准确率为 94.16%，而宽度学习系统的判别准确率为 96.89%。当样本数量为 1000，即约为总样本数的 1/8 时，宽度学习系统分类准确率便高于 99.5%，因此应用宽度学习系统对蘑菇进行是否有毒检测是比较可靠的。同时比较两种算法的训练时间，可以看到宽度学习系统的训练时间要远远小于 BP-神经网络。因此宽度学习系统有着准确率高、训练时间短、判别迅速的特点。

接着通过宽度学习系统的增量学习算法来探究隐藏层节点数 L 与分类准确率的关系。首先，选取其初始特征映射节点 100 个，初始增强节点 50 个。通过宽度学习系统的增量学习算法，每次增加增强节点数 20 个，增加 15 次，直至其达到 330。其在增量过程中的正确率如图 6 及表 4 所示。

由表 4 及图 6 可以看到，随着宽度学习系统隐藏层增强节点数量的增加，蘑菇毒性判别的准确率不断提升，并达到 99.99%。根据宽度学习系统的增量学习算法，若当原系统的性能达不到要求时，可以在原有的宽度学习系统的基础上增加隐藏节点来提升系统的性能，而无需将网络重新进行训练，节省了大量的训练时间，这使得系统实时快速判别蘑菇毒性成为可能。

3 结论

3.1 本文提出了一种基于宽度学习系统的蘑菇毒性判别方法，与传统方法相比，该方法具有准确率高、训练时间短、判别迅速且易拓展的特点，有较强的实用性。宽度学习系统中的大量隐藏层节点保证了该系统能充分学习到输入数据的信息，而由特征节点经非线性变换生成的增强节点则增加了网络中的非线性因素，使得系统能对蘑菇是否有毒进行有效的判别。

3.2 仿真的结果表明，即使在样本量小的情况下，该方法也能保持较高的判别准确率。当样本数量达到一定的数据量后，基于宽度学习系统的蘑菇毒性判别方法的准确率达到了 100%。同时利用宽度学习系统的增量学习算法，使得系统能够快速的更新，使实时判别蘑菇毒性成为可能。与其他机器学习算法相比，该方法理论分析简单且实现容易，易于非机器学习专业的人员学习及使用。

3.3 并且，蘑菇各指标与毒性相关性探究实验表明，蘑菇的帽部特征（形状、表面特性）、气味、各部位（腮、茎、环、菌幕、孢子）颜色是其区分度最大的特征，此结果与人工判别蘑菇毒性所总结的经验是相符的。基于宽度学习系统的蘑菇毒性判别方法消除了人为主观因素的影响，使得判别准确率提升且不受个体经验差异的影响，因此可依据此方法对野外拍摄的蘑菇照片进行识别与毒性判别。

参考文献

[1] 朱元珍,张辉仁,祝英,等.古今毒蘑菇识别方法评价[J].甘肃科学学报,2008,4:40-44

ZHU Yuan-zhen, ZHANG Hui-ren, ZHU Ying, et al. An appraisal of past and present methods for identifying poisonous mushrooms [J]. Journal of Gansu Sciences, 2008, 4: 40-44

（下转第 54 页）