基于 CART 算法对蘑菇毒性的分类与识别

摘　　要

在日常生活中，判断蘑菇是否有毒一般是依靠人们的经验进行的，人们通常认为颜色鲜艳、形状怪异的蘑菇是有毒的，那么这些指标与毒性之间的关系如何，如何应用这些属性指标对蘑菇毒性进行准确判断是一个值得研究方向。本文旨在建立蘑菇分类模型对蘑菇是否有毒进行准确的辨别。

针对问题一：我们应用 CART **算法**对蘑菇的缺失值进行填充，填充的准确率为 76.61%和 89.95%，并将属性指标转换为数值指标实现了数据的数值变换，最后我们运用 SPSS 软件对蘑菇的属性指标进行**主成分分析**，得到了三个主成分。确定了主成分 1 为蘑菇的形态特征，主成分 2 为蘑菇的颜色特征。

针对问题二：首先我们应用 SPSS 软件确定了蘑菇毒性与各项指标具有强相关性，接着我们用 Excel 对其相关性进行可视化，绘制出了条形统计图。利用相关性的分析我们得到了有毒无毒蘑菇在颜色、形状、褶皱上的附生物这三个指标上存在显著差异的结论。

针对问题三：我们基于**神经网络**和 CART 算法建立蘑菇分类模型。同时应用 CART 算法构造出决策树，我们找到了六个影响蘑菇毒性识别的主要因素，并对其重要性进行了排序。

针对问题四：我们从模型的分类准确度、训练分类所用时间这两个角度对模型分类效果进行评估，我们得出的结论是两种模型分类的准确度都大于 90%，模型可靠性强但基于 CART 算法建立的模型要优于基于神经网络建立的模型。之后我们又对模型进行了灵敏度分析，模型的扰动值分别为 0.0011 和 0.0029。

针对问题五：我们运用 SPSS 对毒蘑菇进行**聚类分析**，经过聚类分析，我们将毒蘑菇分为四类，并分析了每一类蘑菇的特征。经过上述问题的分析，我们提出的建议是对毒蘑菇的识别要综合多项指标。

关键词：CART 算法　神经网络　主成分分析　相关性分析　聚类分析

一、问题重述

1.1 问题背景

可食用无毒蘑菇因其营养丰富、味道鲜美，深受广大人民的欢迎。然而，自然界中蘑菇种类繁多，尤其很多有毒蘑菇与可食用无毒蘑菇看起来非常相似，导致每年都有误食毒蘑菇中毒的事件发生，这也是我国食物中毒事件中导致死亡的主要因素之一。民间流传许多鉴别毒蘑菇的方法，简单易行但不能作为鉴别的通用方法。对蘑菇是否有毒的分类与有效识别受到了国内外科研工作者的广泛关注。要求根据附件中的蘑菇数据，建立数学模型研究下列问题。

1.2 问题提出

1、进行数据预处理，如数值规约、属性规约、数据变换、缺失值处理等，并分析蘑菇各指标的基本特征。

2、分析蘑菇的毒性与各指标的相关性，讨论有毒无毒蘑菇在哪些指标上存在显著差异。

3、结合上述分析，构建蘑菇有毒无毒的分类与识别模型，对相关因素的重要性进行排序，确定影响蘑菇毒性识别的主要因素。

4、从影响因素选择、模型方法选择等角度，评估上述所构建判别分类模型在蘑菇有毒无毒方面的识别效果，并开展模型灵敏度和可靠性分析。

5、针对有毒蘑菇进一步进行分类，并分析对应类别下蘑菇的特征；依据所研究成果，给出有毒无毒蘑菇识别的参考建议。

二、问题分析

问题一的分析：问题一要求我们对数据进行预处理。首先我们对数据进行了缺失值处理，删除了干扰属性和缺失值较多的属性后，我们应用 CART 算法填充了附件中的缺失值。接着我们对附件中的数据进行数值变换，即将表格中属性的字母表示换成了相应的数值，实现了字母的数据化。最后，我们对蘑菇的属性进行了主成分分析实现了属性的规约，并根据所得的主成分，分析对应成分的基本特征。

问题二的分析：问题二要求我们分析蘑菇的毒性与各指标的相关性，我们根据问题一对数据的筛选结果，选择显著影响蘑菇毒性的指标进行分析。我们首先用 SPSS 对各个指标进行相关性检验。接着通过计算不同属性不同指标下有毒的概率判断单独指标与毒性之间的相关性，并用 Excel 对研究结果进行可视化。

问题三的分析：问题三是一个分类问题。首先我们应用神经网络，建立蘑菇识别模型。

接着我们又应用 CART 算法基于基尼系数构造决策树建立第二个蘑菇识别模型。同时我们也根据该算法对影响相关指标的重要性进行了排序，并筛选出了影响蘑菇识别的主要因素。

问题四的分析：问题四要求我们对上述模型的识别效果进行评估。我们把附件中的部分

数据作为测试数据集，评估上述模型的识别效果，并基于此结果对模型的可靠性进行分析。之后我们随机改变了蘑菇 500 个属性值，进行模型的灵敏度分析。

问题五的分析：问题五是一个分类问题。我们采用聚类的思想对毒蘑菇进行聚类分析，并指出对应类别下的蘑菇特征。

三、模型假设

1、假设蘑菇的毒性与生长的大小无关；

2、假设蘑菇的毒性不随季节的变化而变化；

四、符号说明

符号	说明
p_i	有（无）毒样本的比例
α	不同指标
$Gini(D_i)$	子集的基尼系数

五、模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 原始数据的缺失值处理

通过观察原始数据我们发现，每一个蘑菇的属性指标都存在缺失值。进一步对各个属性指标进行观察，我们发现有些属性指标较为完整而有些指标缺失较多。我们认为，缺失值较多的指标在填充后会使数据的准确度降低，从而影响对毒蘑菇识别的准确度。故我们删除了 gill-spacing、stem-root、stem-surface、veil-type、veil-color、spore-print-color 这六个缺失值大于 1/5 的指标，对 gill-spacing 和 ring-type 这两组缺失值小于 1/5 的指标进行缺失值填充。

由于每一个蘑菇的指标都存在缺失值，我们无法运用现有数据进行缺失值的填充。我们进入数据来源的网站自己增补了没有缺失值的蘑菇数据。我们用没有缺失值的蘑菇数据对 CART 算法进行训练（CART 算法将在 5.2.3 进行介绍），再将需要填充的数据进行导入。CART 算法对缺失的数据进行填充并对填充结果进行输出。

在填充缺失值的同时，我们也对填充值进行了准确度检验，即让 CART 算法分别对 gill-spacing 和 ring-type 的所有值进行填充，填充后的未缺失值部分与原始数据的未缺失值进行比对，从而得到缺失值填充后的准确度。在进行准确度检验时我们发现蘑菇伞的直径、枝干高度、枝干宽度这三个指标严重影响了缺失值填充的准确度。根据经验与文献查找，我们认为蘑菇的大小指标与蘑菇是否有毒不存在关联，反而会影响研究结果。所以我们剔除了这三个指标，并且接下来的问题研究中也不考虑这三个指标的影响。

我们计算出了两个指标缺失值填充的准确率，如下表所示：

表 1　缺失值填充的准确率

指标	准确率
Gill-attachment	0.7661
Ring-type	0.8995

表 1 所示，缺失值填充的准确率较高，我们认为在误差范围内符合缺失值填充标准。

5.1.2 数据变换

我们发现，表格中的数据均以字母的形式存在，这给接下来的研究带来了巨大的困难。所以我们将 p(有毒)定义为 1，e(无毒)定义为 0，并且应用填充后的数据统计了所有指标有毒蘑菇的占比。依照上述定义可知，有毒占比高的属性越接近 1，无毒占比高的属性越接近 0。根据上述定义，我们对每个指标的字母进行数值变换。我们以其中一个蘑菇为例进行转换，如下表所示：

表 2　数值转换

Class	Cap-shape	Cap-color	Does-bruise-or-bleed	season	Stem-color
p	x	o	f	w	w
1	0.5118	0.7071	0.5594	0.3932	0.4267
habitat	Has-ring	Ring-type	Gill-attachment	Gill-color	
d	t	g	e	w	
0.5439	0.6047	0.4306	0.3750	0.4543	

5.1.3 属性规约

介于蘑菇属性指标繁多，我们希望对相同属性进行合并来减少数据维数，从而提高数据挖掘的效率。我们采用主成分分析法，运用 SPSS 软件对属性指标进行降维。采用 SPSS 统计分析软件对蘑菇 10 个属性指标进行因子分析，经过 KMO 和 Bartlett 检验后，我们得到 KMO 值为 0.749，具有统计学意义。

通过主成分分析，我们得到了三个主成分，并且得到了变量间相关系数矩阵的特征值、方差贡献率、累计方差贡献率，如下表所示：

表 3　主成分列表

	初始特征值			提取平方和载入			旋转平方和载入		
成分	合计	方差%	累积%	合计	方差%	累积%	合计	方差%	累积%
1	3.331	33.312	33.312	3.331	33.312	33.312	3.267	32.670	32.670
2	1.371	13.707	47.019	1.371	13.707	47.019	1.377	13.768	46.438
3	1.109	11.086	58.105	1.109	11.086	58.105	1.167	11.667	58.105
4	0.973	9.734	67.839						
5	0.802	8.024	75.863						

各主成分包含的指标由下表所示：

表 4　主成分包含的指标表

因子数	包含成分名称	方差贡献率	累计方差贡献率
1	cap-shape, does-bruise-or-bleed, has-ring, ring-type, habitat, season	32.67%	32.67%
2	gill-color, stem-color	13.76%	46.43%
3	cap-color, gill-attach	11.66%	58.10%

通过查找相关文献我们认为，主成分 1 主要反应了蘑菇的形态特征，主成分 2 和主成分 3 主要反映了蘑菇的颜色特征。

5.2 问题二模型的建立与求解

5.2.1 相关性分析

我们运用 SPSS 软件粗略的对蘑菇的毒性与其他指标的相关性加以判断。结果表明每个指标与蘑菇的毒性都有显著的相关性。为了更直观的体现各指标与蘑菇毒性的关系，我们求出了每个指标的毒性占比，并用 Excel 进行了可视化。

下图为各指标的毒性占比柱状图。

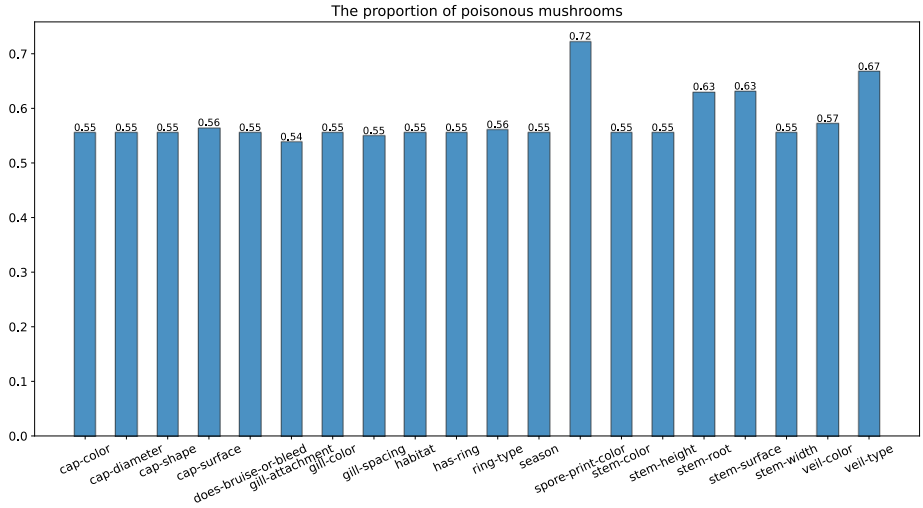

图 1　各指标毒蘑菇占比图

5.2.2 指标分析

由图 3 可以看到，有一些指标与蘑菇是否有毒联系密切。根据图示，我们认为当指标的数值超过 0.6 时，则可以认为这一指标与毒性有明显的相关性。

根据蘑菇伞、蘑菇枝干、蘑菇褶皱的颜色这三张图，我们可以得到蘑菇颜色与蘑菇毒性之间的关系，即 e(红色)、o(橘色)、p(粉色)、r(绿色)、k(黑色)，这五种颜色的蘑菇通常有毒，b(淡黄色)的蘑菇通常无毒，而其他颜色则通常无法进行判别。

蘑菇褶皱上的附生物有时也可以判断蘑菇的毒性，附生物在细毛孔里，大都说明蘑菇无毒，而附生物是直生的，大都说明蘑菇有毒，其他附生物则难以辨别。

蘑菇的形状也可以作为判断蘑菇是否有毒的依据。蘑菇成 b(钟形)或 o(其他形状)则可说

明蘑菇大概率为无毒，成其他形状则无法辨别。

以上分析皆是基于已知数据以及蘑菇的属性指标对蘑菇毒性的辨别。虽然辨别的准确度相对较低，但是却反应出目前人们对毒蘑菇的辨别方式即通过蘑菇的形状、颜色、附生物等对毒蘑菇进行识别。人们普遍认为毒蘑菇的颜色大多美丽艳丽，无毒的蘑菇则以浅褐色、黄色、白色为主。这一观点与上述结论相吻合，但却不是完全有效的辨别方式，接下来我们将用 CART 算法，根据蘑菇的指标特征，对蘑菇进行更为准确的分类。

5.3 问题三模型的建立与求解

传统的毒蘑菇鉴别通常是基于蘑菇的外观特征比如颜色、气味、外形等。但蘑菇种类繁多形态颜色多样难以有效的鉴别是否有毒，所以应用蘑菇形态特征去鉴别蘑菇一直成为困扰科学家的一大难题。本节用神经网络与 CART 算法建立分类模型，通过输入蘑菇的各项属性特征实现对蘑菇是否有毒的自动分类。

5.3.1 基于神经网络建立分类模型

我们基于 LBFGS 算法建立训练神经网络分类模型对蘑菇是否有毒进行分类与识别。神经网络是一种典型的多层前向神经网络，包含输入层、隐含层和输出层。模型的示意图如下：

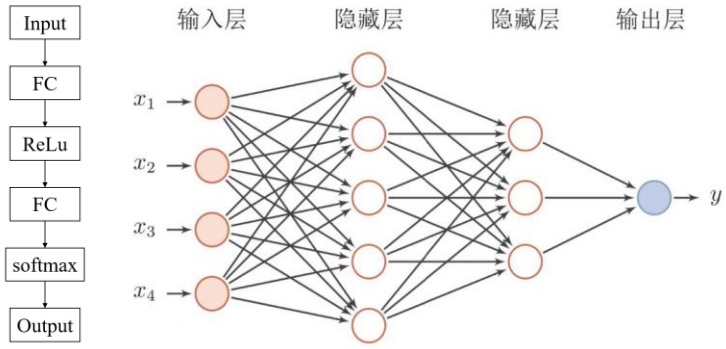

图 2　神经网络示意图

我们将问题一中得到的数据以 8:2 的比例分为训练集和测试集。我们将训练集的数据输入 MATLAB 中，训练该模型使该神经网络可以作为分类器使用，以达到对蘑菇进行分类的目的。

5.3.2 CART 算法

为了提高分类效率以及分类的准确度，我们又使用 CART 算法对蘑菇的毒性重新进行判别。CART 算法是基于基尼系数实现的算法，即选择具有最小基尼系数的值作为分裂属性，并按照节点的分裂属性，采用二元递归的方式形成一棵二叉树。

首先我们需要计算各指标的基尼系数，基尼系数越小说明数据纯度越高。基尼系数计算公式如下：

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

同时我们也需要基于数据中的属性，计算每个属性的基尼系数，计算公式如下：

$$Gini(D,\alpha)=\sum_{i=1}^{m}\frac{|D_i|}{D}Gini(D_i)$$

我们选取基尼系数最小的特征作为分隔点，每次计算时都选取基尼系数最小的作为每层平台的特征。基于上述原理，我们使用 CART 算法建立蘑菇分类模型，并随机抽取表格中80%的数据对该模型进行训练。

5.3.3 指标重要性的排序

应用 CART 算法建立的二叉树，我们实现了对蘑菇 20 个属性的重要性比较。由 CART 算法原理可知，指标的基尼系数越小说明指标的纯度越高，在二叉树当中指标越靠上，也就说明了该指标越重要。我们根据 CART 算法建立的二叉树最终确定了 6 个影响蘑菇毒性识别的主要指标，分别是：stem-color(枝干颜色)、cap-shape(蘑菇伞的形状)、gill-attachment(菌褶上的微生物)、has-ring(是否有圆环)、cap-color(蘑菇伞的颜色)。图 5 显示了基于主要指标判断蘑菇是否有毒的过程。表 5 列出了这些指标重要性的排序。

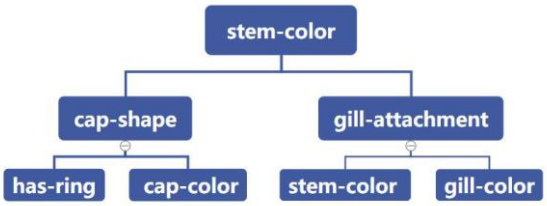

图 3　CART 决策树示意图

表 5　指标重要性排序

重要性	名称
第一重要指标	枝干的颜色
第二重要指标	蘑菇伞的形状、菌褶上的附生物
第三重要指标	是否有圆环、蘑菇伞的颜色、菌褶的颜色

由重要性排序我们可以看出，识别蘑菇是否有毒的最重要因素是蘑菇枝干的颜色。人们通常认为的按照蘑菇伞的颜色是否鲜艳断定蘑菇是否有毒的这种做法虽有一定道理，但蘑菇伞的颜色却不是分辨蘑菇毒性的最重要的指标。而通过蘑菇伞的形状来分别毒蘑菇也具有一定的片面性。

5.4 问题四模型的建立与求解

5.4.1 模型的可靠性分析

我们将剩余 20%的数据作为测试集数据，让两个模型对测试集数据进行分类，用分类结果与真实结果做比较得到了两模型的分类准确率。

用 CART 和神经网络进行蘑菇毒性的分类得到了如下图所示的结果：

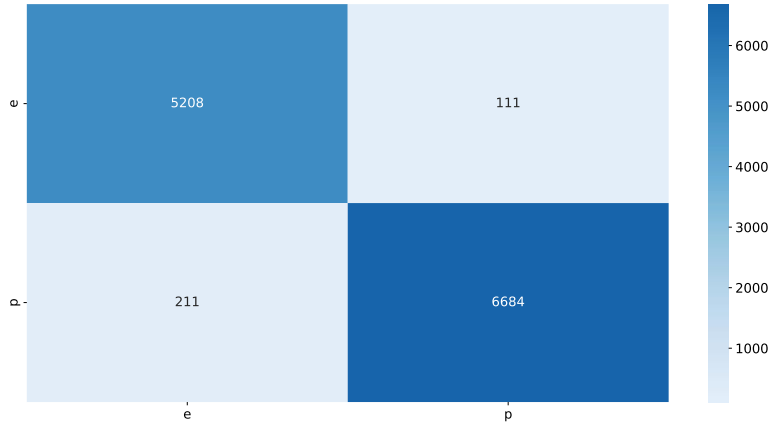

图 4a　CART 分类结果

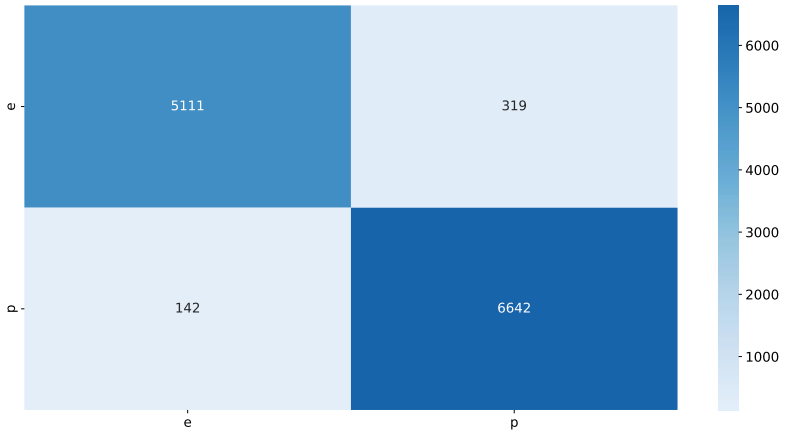

图 4b　神经网络分类结果

用 CART 进行蘑菇分类的准确率为 0.9630，神经网络准确率为 0.9623。从准确率上来看，两模型的准确度都大于 90%，有较高的可靠性，并且 CART 算法建立的模型准确度要优于神经网络模型。

我们又将两模型的训练时间和分类时间进行了比较，发现神经网络模型要训练时间和分类时间要明显长于 CART 模型。并且每次对神经网络进行训练，分类的准确率都会有小幅度的波动，模型的稳健性较低。故我们认为 CART 模型对蘑菇的分类效果要优于神经网络模型。

5.4.2 模型的灵敏度分析

我们随机更改了 500 个蘑菇指标观察分类结果的变化，来检验模型的灵敏度。灵敏度检验结果如下：

表 6　灵敏度检验结果

	基准值	扰动值
CART	0.9702	-0.0029
LBFGS	0.9600	-0.0011

由表格可以发现，模型的扰动值低，可以说明模型的灵敏度低，模型的稳定性强。

5.5 问题五模型的建立与求解

上述模型都是基于蘑菇的各种指标特征对蘑菇的分类，虽然准确度很高但很难应用到实际生活中。下面我们将用聚类分析的方法进一步对毒蘑菇进行分类。

5.5.1 基于聚类分析对毒蘑菇的分类

我们应用 SPSS 对毒蘑菇的数据进行 k-means 聚类。我们把毒蘑菇分为四类，并且设定最大迭代次数为 50 次，实现了对毒蘑菇的分类。分类结果如下：

表 7　k-means 聚类结果

类别	蘑菇个数
1	859
2	315
3	532
4	395

根据上述表格，我们将蘑菇分为四类。根据各类蘑菇的分类特征，我们对每一种蘑菇的特征进行了总结，如下：

表 8　毒蘑菇特征分类

类别	特征
第一类	颜色为白色黄色棕色不易与无毒蘑菇进行区分但蘑菇伞成凹形并且蘑菇表面有伤痕，菌褶上的附生物为直生常见于树叶或草地
第二类	蘑菇伞的颜色为黑色但蘑菇枝干的颜色却为白色或黄色。附着物多附着在细毛孔上或自由附着
第三类	蘑菇伞的颜色为粉红色或绿色，蘑菇枝干没有颜色并且没有菌褶附着物
第四类	蘑菇颜色不易与无毒蘑菇进行区分，但蘑菇伞成球形

5.5.2 蘑菇识别的参考建议

根据问题二到问题五的分析结果，我们对毒蘑菇的识别提出以下建议：

1、判断蘑菇是否有毒时要综合考虑蘑菇的各种指标特征，不能凭借单一指标对蘑菇是否有毒加以辨别。

2、有些毒蘑菇在颜色上无法与无毒蘑菇进行区分，我们建议参考表 8 的分类结果，提高对毒蘑菇识别的准确度。

3、在日常生活中，我们不要贸然食用种类未知的蘑菇。虽然我们已经对蘑菇的分类以及毒蘑菇的特征加以研究，但蘑菇种类繁多，肉眼识别毒蘑菇的准确度有限，所以我们建议在对蘑菇的毒性进行鉴定前，不要食用野生蘑菇。

六、模型的评价

6.1 模型的优点

1、本文应用 CART 算法与神经网络建立蘑菇分类模型，可推广性强。在对其他事物进行分类时，我们仍然可以沿用这两种算法。

2、在对蘑菇进行分类时，我们并不需要将属性指标转换为数值指标，同时我们建立的分类模型需要大量数据先对分类器进行训练，所以不用将蘑菇的数值指标进行删除，提高了建模效率。

3、本文建立的分类模型可以排除缺失值的干扰。如果应用问题一中我们在数据来源网站寻找到的无缺失值的指标对模型进行训练，尽管需要分类的蘑菇存在缺失值，仍可以对蘑菇毒性进行准确分类。

6.2 模型的缺点

需要大量数据对分类器进行训练，不适于小数据的分类问题。

七、参考文献

[1] 王鹏倞.对毒蘑菇毒素的分类与识别探讨[J].科技与创新,2018(11):61-62.DOI:10.15913/j.cnki.kjycx.2018.11.061.

[2]李旺,俞祝良.宽度学习系统在蘑菇毒性判别中的应用[J].现代食品科技,2019, 35(07):267-272+54.DOI:10.13982/j.mfst.1673-9078.2019.7.037.

[3]朱元珍,张辉仁,祝英,蒲凌奎,蒲训.古今毒蘑菇识别方法评价[J].甘肃科学学报, 2008(04):40-44.DOI:10.16468/j.cnki.issn1004-0366.2008.04.035.

[4] 图力古尔,张惠.中国球盖菇科(六):盔孢菌属[J].菌物研究,2012,10(02):72-96.DOI:10.13341/j.jfr.2012.02.003.

[5] 张超群.基于机器学习的毒蘑菇识别与研究[D].山西农业大学,2019.DOI:10.27285/d.cnki.gsxnu.2019.000146.

[6]费芸洁.基于灵敏度分析的神经网络结构优化方法研究[D].苏州大学,2007.[7]张亮,宁芊.CART 决策树的两种改进及应用[J].计算机工程与设计,2015,36(05): 1209-1213.DOI:10.16208/j.issn1000-7024.2015.05.018.