中国农学通报 2015,31(19):232-236
Chinese Agricultural Science Bulletin

基于支持向量机的蘑菇毒性判别研究

樊 帠[1],彭 卫[1],孙 山[2],刘峻呈[3]
([1]四川农业大学商学院,四川都江堰 611830;[2]湖南商学院经济与贸易学院,长沙 410205;
[3]湖南中医药大学中医学院,长沙 410208)

摘 要:毒蘑菇和可食用蘑菇在外表上非常相似,依靠传统方法难以判别。为了实现判别上的自动化和增强可靠性,提出了一种基于支持向量机的蘑菇毒性判别方法。首先给出了数据样本和数据预处理的方法,其次建立C-SVM模型并进行训练,同时依照一对一方法实现了支持向量机的多分类,最后使用定步长探索法获得了模型的最优参数。仿真实验对比分析了不同样本量,不同参数下所提方法的准确度,验证了该方法在蘑菇毒性判别上的可行性。同时,使用神经网络、决策树方法进行分类器间的性能对比,发现与神经网络、决策树的判别结果相比,所提方法具有准确率高、操作方便、实用性强等优点。

关键词:毒蘑菇;支持向量机;分类器;机器学习
中图分类号:S24 **文献标志码**:A **论文编号**:casb15010019

Discriminant Method of Mushroom Toxicity Based on Support Vector Machine

Fan Ge[1], Peng Wei[1], Sun Shan[2], Liu Juncheng[3]
([1]*College of Business, Sichuan Agricultural University*, Dujiangyan Sichuan 611830;
[2]*School of Economics and Trade, Hunan University of Commerce*, Changsha 410205;
[3]*School of Traditional Chinese Medicine, Hunan University of Chinese Medicine*, Changsha 410208)

Abstract: The resemblance between edible mushroom and poisonous mushroom in appearance makes it hard to distinguish them from each other by conventional methods. In order to achieve the automation of judgment and strengthen the reliability, this paper proposed a method to measure the toxicity of mushroom based on support vector machine. To begin with, collection and pre-processing of the sample data were conducted. Then C-SVM model was built up and trained in accordance with one-to-one principle to further achieve multi-classification by support vector machine. At last, constant step length method was applied to obtain the optimum parameters of the model. By comparing accuracy of SVM classification in diverse sample sizes and parameters, the feasibility was verified in simulation experiments. SVM was more accurate, easy-conducting and practical comparing with neural network and decision tree.

Key words: poisonous mushroom; support vector machine (SVM); classifier; machine learning

0 引言

近年来,中国社会经济持续发展,人民生活水平不断提高,食品安全问题却日益凸显。其中蘑菇作为中国人民一种重要食材,在食用安全性上面一直存在很大的争议。人们一方面追求味蕾的享受,希望品尝天然野生蘑菇,另一方面又害怕误食有毒蘑菇。由于中国地广物博,大量有毒蘑菇广泛分布于各个山区乡镇,误食毒蘑菇的事件时有发生。因此,如何找到一种有

基金项目:四川省教育厅基金项目"无源多点定位(MLAT)关键技术研究"(13ZB0287);四川农业大学科研兴趣项目"基于智能手机的实时跟踪及报警系统设计及研究"(2014296)。

第一作者简介:樊帠,男,1992年出生,四川成都人,本科,研究方向为机器学习和推荐系统。通信地址:611830 四川省都江堰市建设路288号 四川农业大学商学院,Email:fange1122@163.com。
通讯作者:彭卫,男,1969年出生,四川安岳人,副教授,硕士生导师,博士,研究方向为智能算法和数据分析与处理。通信地址:611830 四川省都江堰市建设路288号 四川农业大学商学院,Tel:028-87123439,Email:pw7@163.com。

收稿日期:2015-01-05,**修回日期**:2015-04-22。

效的判别蘑菇是否有毒的方法成为了中国专家学者研究的重点[1-3]。

民间判断蘑菇毒性的方法主要是依据其外形特征、颜色特征、气味特征以及分泌物特征等进行判断[1],这些方法对个人经验依耐性很强,且大部分经验只适用于部分地区的蘑菇,判别准确率较低,不具备广泛推广的可行性。学术界内则主要研究毒蘑菇的毒性成分,并分析其中毒机理[2],此类方法虽然准确,但存在着检测成本高、时间花费大、实验条件要求苛刻和难以实现工程应用等缺点。

随着科学技术的发展,机器学习方法成为了人工智能领域的核心[4]。机器学习方法吸取了信息论、概率论、神经科学等多种科学的研究成果,在识别、预测、分类等应用中表现良好[5-7]。从本质上来说,蘑菇毒性的判定问题本身就是一种典型的多维度的、非线性的分类问题,其特点是样本数量少、特征多,这使得传统分类器在解决此类问题时存在着较大的困难。目前尚未见到将机器学习方法应用于蘑菇毒性检测的国内公开文献。

支持向量机(Support Vector Machine,简称SVM)是机器学习中一种较新的方法[8],它基于VC(Vapnik-Chervonenki)维度理论和结构风险最小化原理提出。相比于神经网络、决策树、最小二乘等经典的分类方法,支持向量机在解决非线性、小样本的问题中表现出了特有的优势[7,9]。

笔者提出了一种基于支持向量机的蘑菇毒性判别方法,通过对参数的寻优建立最优的SVM模型,旨在找到一种准确、简便、可靠的蘑菇毒性判别方法应用于工业生产线上蘑菇的自动化检测和识别方面。

1 基于支持向量机的毒性判别方法

1.1 支持向量机简介

支持向量机的最初是从线性可分的思想中提出的,基本思想是寻求2类样本的最优分类线(面),如图1。

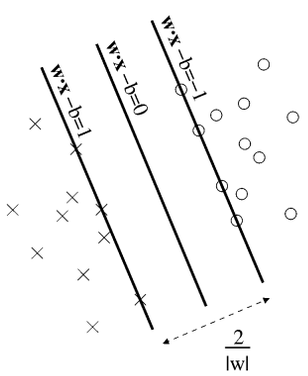

图1 最优分类面

如图1所示,点和叉分别表示2类样本,图中3条线相互平行,其中边缘2条线分别经过2种样本的部分数据,中间这条线被称作最优分类线。可以发现,最优分类线不仅保证了2类数据的分开,并且使得分开的距离最大化。对于高维度维度的情况,最优分类线扩张为一个超平面,使得其余2个超平面的间隔最大。设分类面方程为 $wx-b=0$,于样本集合(x_i, y_i),$i=1\ldots n$,$x\in R^n$,$y\in\{+1,-1\}$,其最优分类面的数学模型见式(1)。

$$\max .Z=\frac{2}{\|w\|} \quad \cdots\cdots (1)$$
$$s.t.\ \ y_i(K(w,x_i)+b)\geqslant 1, i=1,...,n$$

式中:K为内积函数,可以看出,要求2/||w||的最大值即是求||w||的最小值,利用Lagrange最优化方法可以将上述问题转换为其对偶问题,转换后形式见式(2)。

$$\max .Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_jK(x_i,x_j) \quad \cdots\cdots (2)$$
$$s.t.\sum_{i=1}^{n}y_ia_i=0, a_i\geqslant 0, i=1...n$$

其中a_i、a_j分别表示与式(1)中第i个、第j个约束条件的Lagrange乘子,y_i、y_j分别表示第i个、第j个支持向量中y的值。这是一个二次函数的最优化问题,存在唯一的解,并且解中a_i有一部分是不为零的。非零的a_i所对应的样本称为支持向量,解出模型的决策函数如式(3)。

$$f(x)=\mathrm{sgn}(\sum_{i=1}^{n}\alpha_i^*y_iK(x_i,x)+b^*) \quad \cdots\cdots (3)$$

上述内容是线性可分支持向量机的核心思想,然而在实际数据中,大多数情况下数据是线性不可分的。对于数据非线性的问题,通过加入松弛变量来解决,即将式(1)变为式(4)。

$$\max .Z=\frac{2}{\|w\|}+C\sum_{i=1}^{n}\xi_i \quad \cdots\cdots (4)$$
$$s.t.\ y_i(K(w,x_i)+b)\geqslant 1-\xi_i, i=1,...,n$$

在式(4)中$C\geqslant 0$,对于此问题同样可用拉格朗日方法来转化求解,其结果函数和原结果函数相同,只是a_i的范围变为[0,C]。在模型中C实际是作为罚函数的形式存在的,称作惩罚系数,这种模型被称作C-SVM[9]。

另一方面,可以通过一个映射函数将低维度输入空间R^n映射到高维度特征空间H,这样的映射函数在支持向量机中被称作核函数。其中最常使用的是RBF核函数,其公式见式(5)。

$$K(X_i,X_j)=e^{-\gamma\|X_i-X_j\|^2} \quad \cdots\cdots (5)$$

1.2 多分类问题

通常支持向量机的多分类问题的处理方式有2类，一类方法是将多分类问题处理为一系列的二分类问题。另一类是通过改变支持向量机中原始的最优化问题，让其可以适应多分类。笔者选取的方法属于第一类的一对一分类法。假设有n类样本待分类，则每次选择不重复、不相同的2类样本建立一个分类器，共建立$n(n\text{-}1)/2$个分类器。在需要分类时，将样本特征输入所有分类器进行计算，选取被分类次数最多的种类作为最终的输出。

2 实验设计及结果

实验分为2部分。

（1）支持向量机模型仿真，包括数据的预处理、数据的训练与测试、结果分析3部分，实验流程图如图2所示。

（2）与经典分类器的对比实验，为了更全面地分析

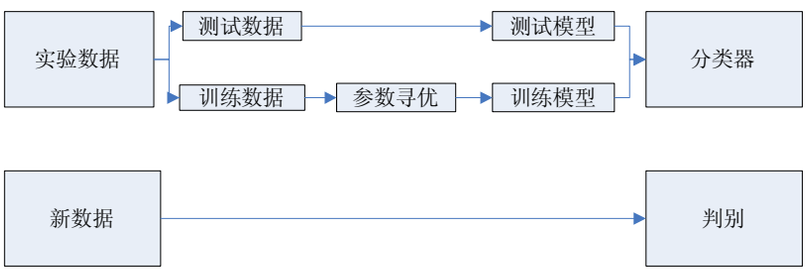

图2 实验流程

笔者所提方法的性能，将实验结果与神经网络、决策树的实验结果进行对比。

2.1 支持向量机模型仿真

使用加州大学欧文分校所提供的蘑菇数据集进行实验[10]，数据共有22个特征，分别为：帽形，帽面，帽色，清肿，气味，鳃—附属，鳃—间隔，鳃—形状，鳃—颜色，茎秆—形状，茎秆—根，茎秆—表面—上环，茎秆—表面—下环，茎秆—色彩—上环，茎秆—色彩—下环，菌幕—数目，菌幕—色彩，环—数目，环—类型，孢子—印记—色彩，数目，生长地。这些特征都可以直接经过观察得到。

数据共有8124组，其中第一列分别用指标{+1,-1}表示蘑菇的毒性，其余各列类似的用1,2…分别表示各样本在各特征下的情况，其部分样本如表1所示。

在Matlab (2013a)环境下进行实验，在Libsvm开源包的基础上进行编程[11]，选取的支持向量机模型为C-SVM模型，选取RBF核函数。通过随机选样，分别选取了4组样本量为400、814、4125、8124的数据进行实验，为保证学习拥有足够的精度，对数据采用mapminmax函数归一化处理。

在完成数据预处理后，实验设计如下：首先，将实验数据按照均匀分布随机分成5份，选取其中1份作为测试数据，其余4份作为训练数据；然后，根据蘑菇数据建立毒蘑菇判别模型，并在测试数据上对蘑菇是否有毒进行预测，统计出预测值的准确率。考虑到实验结果的准确性和避免过拟合，进行5次实验，每次选取不同的测试数据，并将5次实验获得准确率的平均值作为最终准确率。

笔者所提方法需要确定的最重要参数是C和γ，其中C是C-SVM模型即公式(4)中的C，参数γ指的是RBF核函数即公式(5)中的γ。所采用的寻优方法为固定步长和范围的探索方法，并采用交叉训练集预测准

表1 部分蘑菇的部分特征数据

样本编号	毒性	帽形	帽面	帽色	清肿	气味	鳃—附属	鳃—间隔
1	-1	3	4	1	1	8	3	1
2	+1	3	4	10	1	1	3	1
3	+1	1	4	9	1	2	3	1
4	-1	3	3	9	1	8	3	1
5	+1	3	4	4	2	7	3	2
6	+1	3	3	10	1	1	3	1
7	+1	1	4	9	1	1	3	1
8	+1	1	3	9	1	2	3	1

确率。其中,当样本量为400时,参数C和γ寻优过程的绘图如图3(为了更好表示出各参数下准确率的性质,图3中横、纵坐标数据都经过对数化处理,图像以等效果线的形式进行绘制)。

由图3可以看出,在同等样本数量情况下,最差参数和最佳参数的交叉验证准确率的差异超过10%,这表明参数的选取对结果的准确率有重要的影响。

实验分别使用默认参数以及通过寻优得到的参数进行实验,并对比实验结果。表2显示了在不同样本量下,对最优参数C和γ的确定及其准确率结果。

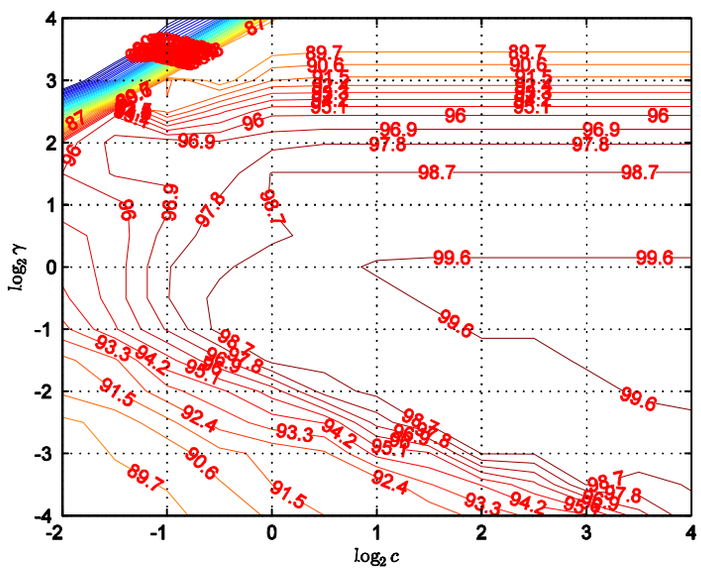

图3 等效果线图

表2 各样本最优参数准确率表

样本量	400个样本	814个样本	4125个样本	8124个样本
参数C	2	1.4142	2	0.3536
参数γ	0.3536	0.7071	2	2.8284
交叉验证准确率/%	98.79	99.84	100	100
默认参数准确率/%	93.18	94.53	95.48	96.31
最优参数准确率/%	98.71	99.87	100	100

由表2可知,随着样本数量的增加,交叉验证准确率、默认参数准确率以及最优参数准确率在不断上升,并且最优化的参数C和参数γ使得该方法在样本数量只有总数量1/10左右时,也能使得判别准确率即最优参数准确率接近100%。这证明了对于毒蘑菇的检测,该方法具有较好的可靠性。

2.2 与BP神经网络及决策树的对比实验

为了更全面地说明问题,应用BP神经网络以及C4.5决策树这2种经典的分类器作为对比[12-16],对比结果见表3。

由表3可知,相比于神经网络、决策树等分类器,基于支持向量机的蘑菇毒性判别方法具有如下优点。

(1)在结果准确率上有着较大优势。相同样本量下,基于支持向量机的蘑菇毒性判别方法的判别精确度均高于其他分类器。例如,在814个样本的这组实验中,BP神经网络的判别准确率只有95.36%,C4.5决策树算法的判别准确率有96.88%,而对基于支持向量机的方法,其判别准确率达到了99.87%。

表3 3种分类器准确率对比表 %

模型	400个样本	814个样本	4125个样本	8124个样本
支持向量机	98.71	99.87	100	100
BP神经网络	95.18	95.36	95.48	95.55
C4.5决策树	96.29	96.88	97.30	97.41

(2)对数据量变换更敏感。随着数据量的增加，BP神经网络与C4.5决策树模型判别的准确率上升缓慢，特别是神经网络方法，数据对其准确率的提升几乎停滞，而同等样本情况下，基于支持向量机的方法的准确率随着数据量增大而提升并达到100%。

基于此，相比于经典分类器，基于支持向量机的方法更适用于蘑菇毒性的判别。

3 结论

笔者提出了一种基于支持向量机的蘑菇毒性判别方法，与传统方法相比，基于支持向量机的蘑菇毒性判别方法具有准确性高、成本低、易于推广等特点，具有较强的实用性。

基于支持向量机的判别方法可作为工厂中自动化蘑菇检测处理流程的理论基础；也可以依据该方法编制手机程序，对拍摄的野外蘑菇图片进行自动识别和毒性检测。

在选取最优模型参数时，笔者所用的是定步长的二维探索法，该方法在某些特殊情况下的收敛速度较慢。如何结合非线性最优方法(例如遗传算法、粒子群算法等)来提高寻优速度，是下一步研究的重点。

4 讨论

笔者基于支持向量机，提出了一种可自动化的蘑菇毒性判别方法。仿真实验结果表明，所提出的基于支持向量机的蘑菇毒性判别方法即使在训练样本量较小时，判别准确率也可接近100%。

与传统的人工蘑菇毒性判别方法相比[1]，虽然都是依据蘑菇的外形、颜色等因素来判定蘑菇的毒性，但由于基于支持向量机的方法排除了人为主观因素，在判别准确率上有较大提升。与经典的神经网络和决策树分类器相比[12-15]，笔者所提方法在准确率、数据灵敏度上有较强的优势。这是由支持向量机的特性引起的，支持向量机的核心在于：(1)最大化支持向量的间隔；(2)核函数。第一点保证了支持向量机分类的最优性；第二点保证了支持向量机可应用于多分类问题。相比于经典的神经网络和决策树等分类器，这2点支撑了支持向量机在小样本、非线性问题中的优势。

当实验样本达到一定数据量后，基于支持向量机的方法判别准确率达到了100%，当实验样本达到一定数据量后，基于支持向量机的方法判别准确率达到了100%，这是因为支持向量机在解决小样本、非线性的问题时具有一定优势。在实际中，蘑菇数据具有小样本、非线性的特点，所以基于支持向量机蘑菇毒性判别方法在实际工程中是适用的。

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,2008,20(4): 40-44.

[2] 包海鹰,图力古尔,李玉.蘑菇的毒性成分及其应用研究现状[J].吉林农业大学学报,1999,21(4):107-113.

[3] 图力古尔,包海鹰,李玉.中国毒蘑菇名录[J].菌物学报,2014,33(3): 517-548.

[4] Drew M S, Li Z N, Zhong X. Video dissolve and wipe detection via spatio-temporal images of chromatic histogram differences[A]. In: 2000 International Conference on Image Processing,2000. Proceedings[C]. Piscataway: IEEE press,2000,3:929-932.

[5] 徐忠杰,杨永国,汤琳.神经网络在矿井水源判别中的应用[J].煤矿安全,2007(2):4-7.

[6] 王珏,石纯一.机器学习研究[J].广西师范大学学报:自然科学版,2003,21(2):1-15.

[7] 陈果,周伽.小样本数据的支持向量机回归模型参数及预测区间研究[J].计量学报,2008,1(1):92-96.

[8] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995,20(3):273-297.

[9] Tay F E H, Lijuan Cao. Application of support vector machines in financial time series forecasting[J]. Omega,2001,29(4):309-317.

[10] Bache K, Lichman M. UCI Machine Learning Repository[EB/OL]. Http://archive.ics.uci.edu/ml,1987-04-27/2014-12-27.

[11] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a library for support vector machines[J].ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27.

[12] 魏海坤编著.神经网络结构设计的理论与方法[M].北京:国防工业出版社,2005:20-43.

[13] 沈花玉.神经网络在医学诊断中的应用研究[D].天津:天津理工大学,2007:22-26.

[14] 刘彩红.BP神经网络学习算法的研究[D].重庆:重庆师范大学,2008:24-30.

[15] 郭炜星.数据挖掘分类算法研究[D].杭州:浙江大学,2008:19-26.

[16] 尹阿东,郭秀颖,宫雨,等.增量决策树算法研究[J].微机发展,2005,15(2):63-65.