基于数据挖掘的触诊成像乳腺癌智能诊断模型和方法

张旭东[1]，孙圣力[1]，王洪超[2]
1. 北京大学软件与微电子学院，北京 100089；
2. 北京先通康桥医药科技有限公司，北京 101300

摘要
为了辅助医护人员利用触诊成像技术判定乳腺癌，提出了触诊成像乳腺癌智能诊断模型和方法。采用乳腺癌早期筛查及风险评估的临床数据，以触诊成像诊断结果为对比数据，通过决策树等机器学习算法以及投票法，对乳腺肿瘤的良恶性质进行判定。使用SMOTE算法对数据进行处理，建立了诊断模型和方法，自动完成对乳腺肿瘤性质的诊断。实验结果表明，乳腺癌正确筛查的准确性达到98%，提出的方法具有很好的应用价值。

关键词
智能诊断；临床数据；机器学习；SMOTE算法

中图分类号：TP18　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2019005

Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining

ZHANG Xudong[1], SUN Shengli[1], WANG Hongchao[2]
1. School of Software & Microelectronics, Peking University, Beijing 100089, China
2. Sinotau Pharmaceutical Group, Beijing 101300, China

Abstract
In order to assist the medical staff to diagnose breast cancer more effectively by palpation imaging technology, intelligent diagnosis model and method of palpation imaging breast cancer were established. Based on clinical data for early breast cancer screening and risk assessment, machine learning algorithms of decision tree, neural network, SVM, logistic regression, Bayesian network and five voting methods were adopted to distinguish breast tumor, or positive and negative outcome in algorithms. The positive sample data was incremented by the SMOTE algorithm, intelligent diagnosis model was established, and model can automatically diagnose breast tumors. Palpation imaging intelligent diagnosis model of breast cancer correctly screens all cases of breast cancer confirmed by pathology, and the accuracy of the model is as high as 98%. The intelligent diagnosis model is excellent as a screening modality for the detection of breast cancer.

Key words
intelligent diagnosis, clinical data, machine learning, SMOTE algorithm

1 引言

近年来，乳腺癌已成为威胁女性健康的恶性疾病，发病年龄集中于45~55岁，发病率则随着年龄的增长呈上升态势。提高广大妇女的乳腺健康意识，加强和规范乳腺癌筛查工作，以便早诊早治，对于降低乳腺癌死亡率至关重要。在乳腺癌筛查中应以较少的人力、物力取得较大的社会效益，即选择灵敏、经济的检测手段，制定最佳的筛查方案[1-3]。

鉴于触诊成像在大规模人群筛查中体现出的快速高效的独特优势[4-7]，本文结合机器学习相关技术，采用乳腺触诊诊断仪收集的临床数据，进行诊断模型训练，以乳腺癌临床病理诊断结果为判读标准，建立了一套基于触诊成像的乳腺癌智能诊断方法，以实现乳腺癌的智能化判定，进而提高大规模人群乳腺癌筛查的效率。

在医疗领域，大数据的取得及应用至关重要。大部分数据是通过文献、临床数据、结构化数据、非结构化数据及第三方数据库等渠道获取的。医疗数据存在以下特点和问题。首先，医疗数据具有显著的特殊性及复杂性，要在短时间内积累大量有价值的数据，难度和成本很高；其次，医疗数据往往面临不平衡数据集的问题，样本种类不平衡会导致整个数据集难以有效地运用，数据无法发挥其最大效能。因此，从不同渠道获取数据后，应进行数据清洗，确保数据质量，并在数据转换、重新建构后，将数据存入数据库以供使用。医疗智能诊断旨在帮助医疗机构或医生个人利用信息技术对医学数据进行收集、管理及分析。本文通过积累相关医学知识，利用数据清洗、数据增强等方式提升数据的价值，并运用相关机器学习算法进行乳腺癌预测，建立了一套触诊成像乳腺癌智能诊断方法。

2 乳腺癌智能诊断建模流程

笔者参与的乳腺触诊成像健康体检人群乳腺癌早期筛查研究项目积累了多家医院的临床数据。本文基于这些数据，以触诊成像诊断结果为对比数据，进行相关的预测研究。所有触诊成像被诊断为乳腺癌的阳性标本均经过病理诊断验证，在乳腺癌样本数据中随机选择3个数据集（分别表示为数据集1、数据集2、数据集3），数据量分别为13 428条、1 554条、902条，总计15 884条数据样本。

综合考虑各方面因素和临床数据的特点，运用机器学习中常用的决策树、神经网络、支持向量机（support vector machine，SVM）、逻辑回归及贝叶斯网络5种算法[8-12]，再结合多种投票法，进行乳腺肿瘤的形态预测和判定。

数据在经过预处理等相关操作后，运用合成少数类过采样技术（synthetic minority over-sampling technique，SMOTE）[13-14]，将阳性样本进行合理范围的增量，以解决不平衡数据集问题。对模型进行测试及改进，选择最佳分类模型和方法，并综合利用准确率、召回率等指标，评估分类模型的优劣，得到高质量的乳腺诊断模型，提升总体辅助诊断水平。

整个建模流程如**图1**所示。

3 数据清洗与准备

依据数据清洗（data cleaning）[1]的原则，按**图2**所示过程进行数据清洗。

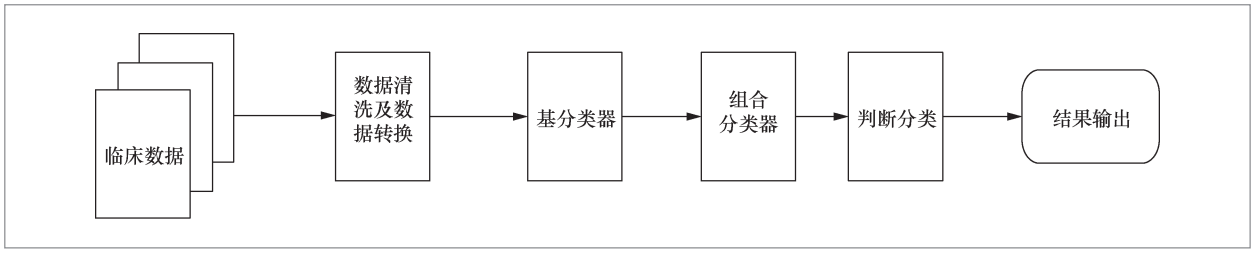

图1 乳腺癌智能诊断建模流程

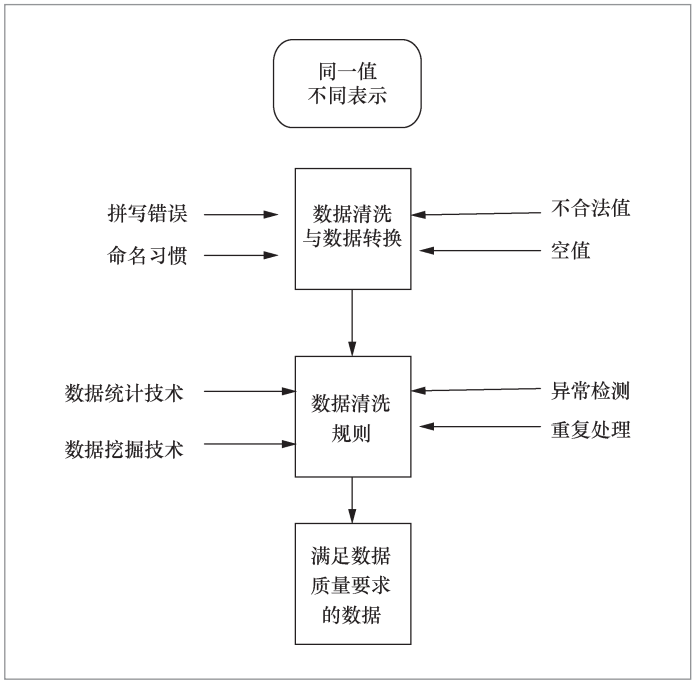

图2 数据清洗流程

原始临床数据有位置、象限、压力值、肋骨干扰、3D峰值、2D颜色、3D峰顶形状、3D形状、3D基底、3D动态、2D形状、2D动态颜色分布、血流灌注指数（PI）诊断结果及病理结果14个属性。其中，压力值及肋骨干扰两个属性对智能诊断系统并无显著影响，故而剔除。为确保数据的完整性，将36个含有缺失值及62个含有噪音值的数据样本剔除。各属性数据缺失量与噪音数据量如**图3**所示。

整个数据集内初始的阳性样本有135个，占所有数据的0.85%。由于阳性数据与阴性数据的比例极不平衡，故而进行了样本数据的整理。在数据查重时，发现排除位置及象限两个不影响结果的属性后，有168个阴性数据与阳性数据属性相同。为避免错失恶性病例情况的发生，将这168个原本

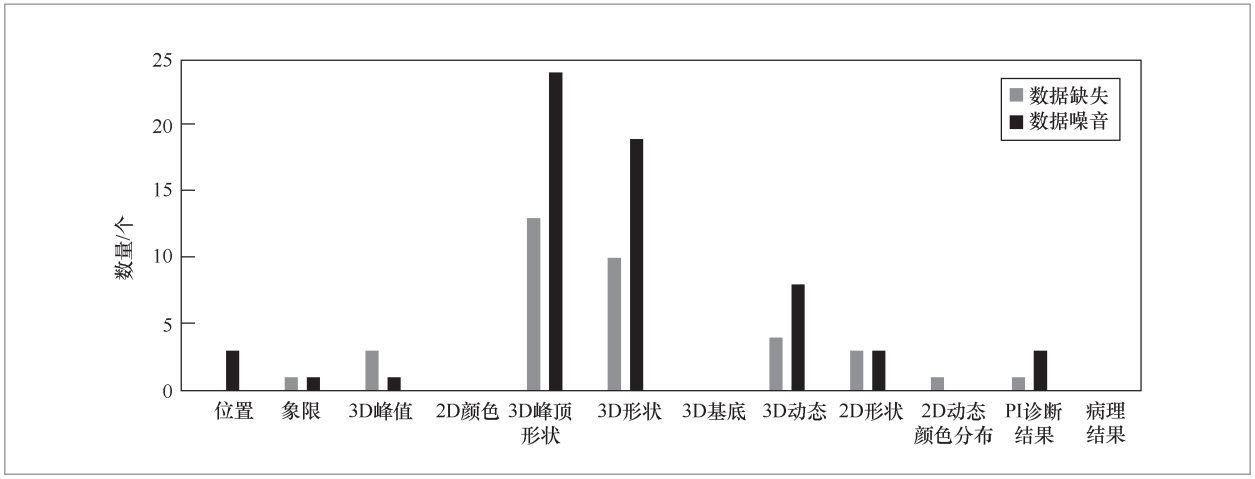

图3 各属性数据缺失量与噪音数据量

标为阴性而实则为阳性的数据样本更改成阳性，以提高数据的准确性。查重前后阳性数据数量见**表1**。

SMOTE算法[2]通过采样操作解决类别间比例相差悬殊的问题。当数据集类别不均衡[2]时，一般采取随机欠采样和随机过采样两种方式来处理。本研究中抽取新值的SMOTE算法示意如**图4**所示，依次遍历数据集中每个集合，直到处理完所有数据为止。最后，将新增加点的集合加至原有数据集的恶性病例类别中，并构成新的数据集。该算法避免了随机过采样复制样本带来的样本数据不准确的问题，解决了模型学习到的信息过于特别而不够泛化的问题。

本研究依据混淆矩阵的分类指标[2]进行模型定量评估，包括准确率（accuracy）、精确度（precision）、召回率（recall）、真阳性率（true positive rate）、F值，其中，召回率又被称为灵敏度（sensitivity）。机器学习中常用准确率与召回率作为参考指标，各指标定义如下：TP为将阳性样本预测为阳性样本的样本数，FN为将阳性样本预测为阴性样本的样本数，FP为将阴性样本预测为阳性样本的样本数，TN为将阴性样本预测为阴性样本的样本数。准确率（正确率）=(TP+TN)/总样本数，精确率=TP/(TP+FP)，召回率=TP/(TP+FN)，F值=正确率×召回率×2/(正确率+召回率)。

本文使用SMOTE算法进行阳性样本增量，数据总量为15 790条（阳性数据303条），使用SMOTE算法第1次和第2次处理数据后，阴性样本与阳性样本的比例分别为25:1和13:1，比例依然不平衡。使用SMOTE算法第3次处理数据后，阳性数据增加至2 424条，阴性样本和阳性样本比例约为6:1，数据集的样本种类较先前数据集更合理且平衡。详细结果见**表2**。

通过使用SMOTE算法3次处理数据后，近邻点K值在1到7中选择并比较结果。近邻点K值是SMOTE算法中生成新样本的参数。经过对比发现，K=7时呈现过拟合的现象，即分类结果有明显下

表1 查重前后阳性数据

	数据集1	数据集2	数据集3
查重前阳性数据量/条	50	11	74
查重后阳性数据量/条	185	22	96

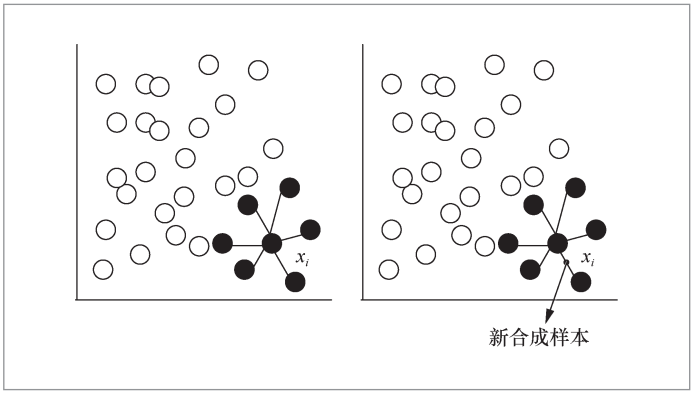

图4 SMOTE算法示意

表2 使用SMOTE算法对阳性数据增量后训练集的结果

数据状况	使用SMOTE次数/次	阴性数据量/条	阳性数据量/条	准确率	精确率	召回率	F值
训练集A	1	15 487	606	99.21%	90.53%	88.03%	89.26%
训练集B	2	15 487	1 212	98.72%	89.80%	92.71%	91.23%
训练集C	3	15 487	2 424	98.75%	92.30%	98.99%	95.53%

滑的趋势，故选择结果表现较优秀的值，即K=6值。

4 模型训练与预测

随后进行数据集抽取。乳腺癌分类属二元分类问题，故将数据内容定义为标准型数值{N,P}，符合数据集要求。在实验设计过程中，将数据集分为训练集及测试集两部分。首先从数据清洗及查重后的数据集内抽取90%的数据作为训练集；在经数据清洗后的数据集内，随机抽取6份数据组成测试集（A~F），每份抽取10%的数据样本，特殊测试集1、特殊测试集2由两份单独的数据集组成，进行最后的模型评估。数据分布见**表3**。

本文选用决策树、神经网络、SVM、逻辑回归、贝叶斯网络作为基分类器，基于训练集进行模型训练。然后，基于上述分类模型进行乳腺癌预测，并根据预测结果进行模型筛选和优化。各基分类器具体预测结果见**表4**。

在基分类器参数最优的前提下，本文将神经网络[15]由原来的单一隐含层调整为两个隐含层，提高了网络的分类能力。结构优化前后的结果对比见**表5**。

在上述基分类器模型预测的基础上，

表 3 数据集的分配状况及说明

数据名	数据状况	数据说明
训练集	16 120条数据	数据经过预处理、查重并处理、SMOTE阳性增量处理
特殊测试集1	1 549条数据	数据经过预处理
特殊测试集2	902条数据	数据经过预处理
随机测试集A~F	每个数据集有1 579条数据	数据经过预处理

表 4 基分类器训练和测试结果

算法名	准确率	精确率	召回率	F值
决策树	99.59%	89.66%	61.42%	72.90%
SVM	99.75%	92.30%	98.99%	95.53%
	98.67%	99.59%	99.59%	99.59%
	98.67%	98.75%	98.75%	98.75%
贝叶斯网络	93.21%	68.86%	90.55%	78.23%
神经网络	98.57%	95.41%	92.70%	94.63%
	96.57%	94.89%	93.87%	93.99%
逻辑回归	93.98%	81.11%	72.06%	76.31%

表 5 数据集的分配状况及说明

数据状况	参数调整	模型调整	基分类器	准确率	精确率	召回率	F值
训练集	采用	不采用	神经网络	98.57%	95.41%	93.87%	94.63%
训练集	采用	采用	神经网络	99.15%	95.16%	98.75%	96.92%

再进行预测算法和模型的优化选择。

组合分类技术是最主要的提高分类器精确度的方法。将通过多个分类器得出的结果作为最终判断的依据，从而避免单一分类器产生的判断误差或片面性信息，以优化分类效果。本文提出的乳腺癌组合预测诊断方法的处理流程如**图5**所示。

通过上述实验，笔者发现决策树、SVM及神经网络3种算法在乳腺癌智能诊断系统中呈现较好的结果，故将贝叶斯网络及逻辑回归[12,16]两种算法剔除，仅保留决策树、SVM及神经网络3种算法。由**表6**可以得知，仅以3种算法作为模型，其准确率及精确率都有显著提升。

在以3种算法作为模型的基础上，笔者构建了一种基于投票选择的组合预测优化方法。在本文乳腺癌预测诊断的方法中，设计了4种投票组合法，包含一票确定法、两票确定法、多票确定法和加权投票法A。因优化的模型中只有3种算法，多票确定法与两票确定法的结果相同，所以删除了多票确定法。

表7结果显示，在优先保证召回率的前提下，加权投票法A与一票确定法结果相同。综合考虑先前阶段的实验对比，选择加权投票法A作为优化后模型的投票方法。

5 实验结果和分析

将使用SMOTE算法的次数设为3、近邻点K设为6，采用3种算法（决策树、SVM、神经网络）及加权投票法A进行最终的训练并建模。随机测试集A~F、特殊数据集1和特殊数据集2对训练集模型验证的结果见**表8**。

图6的结果是8份测试集的平均结果，包括召回率、精确率、准确度及F值4项结果。**表8**结果显示，8份测试集的准确率达97%，说明模型对数据的判断能力很高。此外，随机测试集A~F、特殊数据集1和特

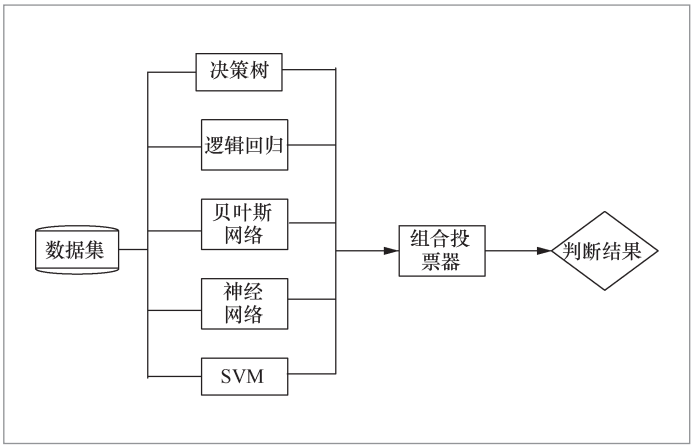

图 5 乳腺癌组合预测诊断方法流程

表 6 不同模型的比较结果

数据状况	模型概况	准确率	精确率	召回率	F值
训练集	以决策树、SVM、神经网络、贝叶斯网络及逻辑回归5种算法作为模型	97.38%	83.71%	100%	91.13%
训练集	以决策树、SVM和神经网络3种算法作为模型	98.64%	90.84%	100%	95.20%

表 7 不同投票器的比较结果

数据状况	投票器	准确率	精确率	召回率	F值
训练集	一票确定法	98.64%	90.84%	100%	95.20%
训练集	加权投票法A	98.64%	90.84%	100%	95.20%
训练集	两票确定法	99.41%	96.47%	99.26%	97.85%

表 8 测试集对训练集模型的测试结果

数据名称	准确率	精确率	召回率	F值
随机测试集A	97.15%	15.09%	100%	26.23%
随机测试集B	97.72%	35.71%	100%	52.63%
随机测试集C	97.15%	16.67%	100%	28.57%
随机测试集D	97.78%	33.96%	100%	50.70%
随机测试集E	97.21%	18.52%	100%	31.25%
随机测试集F	98.16%	32.56%	100%	49.12%
特殊数据集1	94.96%	48.68%	100%	65.49%
特殊数据集2	97.15%	30.56%	100%	46.81%

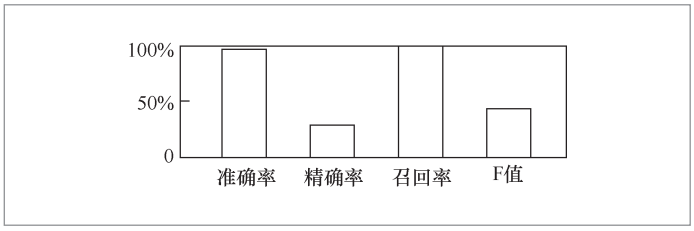

图 6 测试集平均结果

殊数据集2的召回率皆达100%，即所有阳性样本都能被正确地判断出来，说明预测方法的判断结果具有良好的临床辅助诊断应用价值。

6 结束语

本文构建了基于触诊成像的乳腺癌智能诊断模型，给出了5种主要分类算法，通过数据预处理、样本调优等操作，整理出训练和测试数据集。在此数据集的基础上，抽取训练集与测试集，通过训练集训练，建立分类模型及组合投票器，最终判断结果。基于前期预备与调研工作，在保证数据质量的前提下，运用特殊数据进行实验，最终结果在召回率与准确度指标上表现优异。鉴于医疗诊断模型的结果关系重大，笔者后续将持续追加新数据组成新的数据集，不断对模型进行训练，使模型更加完善，以期提供更加高效的临床诊断工具。

参考文献:

[1] 郝淳敏. 乳腺癌二级预防筛查模式的研究[J]. 中华健康管理学杂志, 2012, 6(2): 131-132.
HAO C M. Research on secondary preventive screening model for breast cancer[J]. Chinese Journal of Health Management, 2012, 6(2): 131-132.

[2] 毕晓峰, 徐志坚. 乳腺癌筛查技术研究现状与展望[J]. 中华健康管理学杂志, 2016(4): 321-325.
BI X F, XU Z J. Technology of breast cancer screening: current situation and prospect[J]. Chinese Journal of Health Management, 2016(4): 321-325.

[3] 王本忠, 付君. 乳腺触诊成像影像诊断学[M]. 北京: 科学出版社, 2016.
WANG B Z, FU J. Diagnostics of breast palpation imaging[M]. Beijing: China Science Publishing House, 2016.

[4] 宋颖, 李静, 张仁知, 等. 乳腺可视化触诊成像系统鉴别诊断乳腺良恶性病变: 与临床触诊、X线及超声对照[J]. 中国医学影像技术, 2014, 30(4): 527-530.
SONG Y, LI J, ZHANG R Z, et al. Palpation

imaging system in differentiating benign and malignant breast lesions: comparedwith physical examination, mammographyand ultrasonograph[J]. Chinese Journal of Medical Imaging Technology, 2014, 30(4): 527-530.

[5] 盖小荣, 王振捷, 王健, 等. 触诊成像系统在乳腺疾病健康筛查中的应用研究[J]. 中华健康管理学杂志, 2013, 7(6): 398-401.
GAI X R, WANG Z J, WANG J, et al. The application of palpation imaging in the screening of breast disease[J]. Chinese Journal of Health Management, 2013, 7(6): 398-401.

[6] 于志勇, 左文述, 刘岩松, 等. 乳腺触觉成像诊断系统的临床应用评估[J]. 中华肿瘤防治杂志, 2011(1): 50-53.
YU Z Y, ZUO W S, LIU Y S, et al. Clinical evaluation of breast palpation imaging system[J]. Chinese Journal of Cancer Prevention and Treatment, 2011(1): 50-53.

[7] 方礼妮, 郑春兰. 女性乳腺癌患者主要照顾者抑郁状况及影响因素研究[J]. 中华健康管理学杂志, 2015, 9(3): 209-211.
FANG L N, ZHENG C L. Study on depression status and influencing factors of main caregivers of female breast cancer patients[J]. Chinese Journal of Health Management, 2015, 9(3): 209-211.

[8] 王忠民, 曹洪江, 范琳. 一种基于卷积神经网络深度学习的人体行为识别方法[J]. 计算机科学, 2016, 43(11A): 56-58, 87.
WANG Z M, CAO H J, FAN L. Method on human activity recognition based on convolutional neural networks[J]. Computer Science, 2016, 43(11A): 56-58, 87.

[9] 沈花玉, 王兆霞, 高成耀, 等. BP神经网络隐含层单元数的确定[J]. 天津理工大学学报, 2008, 24(5): 13-15.
SHEN H Y, WANG Z X, GAO C Y, et al. Determining the number of BP neural network hidden layer units[J]. Journal of Tianjin University of Technology, 2008, 24(5): 13-15.

[10] 郑伟, 马楠. 一种改进的决策树后剪枝算法[J]. 计算机与数字工程, 2015(6): 960-966, 971.
ZHENG W, MA N. An improved post-pruning algorithm for decision tree[J]. Computer and Digital Engineering, 2015(6): 960-966, 971.

[11] 张琪, 周琳, 陈亮, 等. 决策树模型用于结核病治疗方案的分类和预判[J]. 中华疾病控制杂志, 2015(5): 510-513.
ZHANG Q, ZHOU L, CHEN L, et al. A decision tree model for classification and prediction of tuberculosis treatment[J]. Chinese Journal of Disease Control & Prevention, 2015(5): 510-513.

[12] 李雅秋, 王旗. 构建用于预测中药化学成分心脏毒性的定量构效关系模型[J]. 北京大学学报(医学版), 2017, 49(3): 551-556.
LI Y Q, WANG Q. Quantitative structure-activity relationship model for prediction of cardiotoxicity of chemical components in traditional Chinese medicines[J]. Journal of Peking University Health Science, 2017, 49(3): 551-556.

[13] 徐斌, 苏一丹, 黄山. 基于km-smote和随机森林的不平衡数据分类[J]. 计算机技术与发展, 2015, 25(9): 17-21.
XU B, SU Y D, HUANG S. Classification of imbalance data based on km-smote algorithm and random forest[J]. Computer Technology and Development, 2015, 25(9): 17-21.

[14] 李克文, 杨磊, 刘文英, 等. 基于RSBoost算法的不平衡数据分类方法[J]. 计算机科学, 2015, 42(9): 249-252, 267.
LI K W, YANG L, LIU W Y, et al. Classification of imbalance data based on KM-SMOTE algorithm and Random Forest[J]. Computer Technology and Development, 2015, 42(9): 249-252, 267.

[15] 肖魏娜, 张为群, 王玲玲. 一种基于BP神经网络的软件需求分析风险评估模型的研究[J]. 计算机科学, 2011, 38(4): 199-202.
XIAO W N, ZHANG W Q, WANG L L. Study on evaluation model of software requirement analysis risk based on neural network[J]. Computer Science, 2011, 38(4): 199-202.

[16] 张月, 黄钢, 章小雷, 等. 贝叶斯网络在医

学领域中的应用研究[J]. 中国医学创新, 2013(4): 145-146.
ZHANG Y, HUANG G, ZHANG X L, et al. Study on application of bayesian networks in the medical field[J]. Medical Innovation of China, 2013(4): 145-146.

作者简介

张旭东(1991-), 男, 北京大学软件与微电子学院硕士生, 主要研究方向为深度学习、计算机视觉等。

孙圣力(1979-), 男, 北京大学软件与微电子学院副教授, 主要研究方向为大数据管理、数据挖掘、图数据库、智慧医疗等。

王洪超(1968-), 男, 就职于北京先通康桥医药科技有限公司, 主要研究方向为乳腺触诊成像技术的开发和临床应用研究。

收稿日期: 2018-11-06
通信作者: 孙圣力, slsun@ss.pku.edu.cn
基金项目: 江苏省自然科学基金资助项目(No.BK20151132)
Foundation Item: The Natural Science Foundation Item of Jiangsu Province(No.BK20151132)