Journal of Computer Applications
计算机应用,2015,35(10):2793 -2797
ISSN 1001-9081
CODEN JYIIDU
2015-10-10
http://www.joca.cn

文章编号:1001-9081(2015)10-2793-05
doi:10.11772/j.issn.1001-9081.2015.10.2793

基于属性约简和支持向量机集成的乳腺癌诊断决策

卢星凝[1],张　莉[1,2*]

(1.苏州大学 计算机科学与技术学院,江苏 苏州 215006;　2.江苏省计算机信息处理技术重点实验室(苏州大学),江苏 苏州 215006)

(*通信作者电子邮箱 zhangliml@suda.edu.cn)

摘　要:针对遗传算法(GA)与支持向量机(SVM)集成相结合的疾病诊断方法存在属性冗余的问题,提出了一种改进的约简和诊断乳腺癌决策方法。该方法将最小化约简属性个数、最大化区分矩阵可区别属性的个数以及最大化约简属性对决策属性的依赖度这三种目标函数相结合作为GA的适应度函数。在约简属性后取多个子集,以便利用SVM集成学习。在UCI数据库中乳腺癌数据集的实验表明,与原始的SVM算法相比,该方法在分类诊断的准确度以及敏感性方面有一定的提高,其中分类准确度至少提高了2%。

关键词:粗糙集;支持向量机;属性约简;乳腺癌诊断;遗传算法

中图分类号:TP391.4　　**文献标志码**:A

Diagnosis decision of breast cancer combining with attribute reduction and support vector machine

LU Xingning[1], ZHANG Li[1,2*]

(1. *School of Computer Science and Technology, Soochow University, Suzhou Jiangsu* 215006, *China*;
2. *Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou Jiangsu* 215006, *China*)

Abstract: In the disease diagnosis approach of combining with Gene Algorithm (GA) and Support Vector Machine (SVM) ensemble, the attribute redundancy problem still exists. A decision method for diagnosis of breast cancer was proposed based on attribute reduction and SVM. The proposed attribute reduction method took minimizing the attribute number, maximizing the difference attribute number in discernibility matrix and maximizing the dependency degree of condition reduction attributes on decision attributes as the fitness function of GA. After attribute reduction, multiple attribute subsets were selected for SVM ensemble learning. Compared with SVM, experimental results on the breast cancer dataset from UCI databases validate that the classification accuracy increases by 2 percent at least.

Key words: Rough Set (RS); Support Vector Machine (SVM); attribute reduction; breast cancer diagnosis; Genetic Algorithm (GA)

0　引言

近年来,癌症成为威胁人类健康的主要杀手之一。据美国癌症学会(American Chemical Society, ACS)2014年公布的数据报告显示[1],乳腺癌仍是女性中最常见的癌症疾病(41%),其次是子宫颈癌(8%)和结肠癌(8%)。研究发现,乳腺癌的发病存在一定的规律性,即具有与乳腺癌发病有关的各种危险因素的女性更容易患病。研究者们希望通过对这些危险因素的分析,达到对乳腺癌的早发现与早诊断,从而争取治疗时间,提高治疗效果,减小乳腺癌对生命威胁的几率。因而,做好乳腺癌疾病的诊断预防工作十分关键[2]。

乳腺癌传统的医学诊断方法主要有:X射线诊断、CT扫描、细针穿刺细胞病理学检查、超声波或核磁共振成像等。其中,细针穿刺细胞病理学检查(Fine-Needle Aspiration, FNA)不仅可以明确诊断乳房肿块,并且可以最简单、便捷地对乳腺肿瘤进行检查[3]。

来自Wisconsin大学的研究团队率先对FNA进行诊断工作[4],提出了基于机器学习技术的乳腺癌辅助诊断,也称智能诊断,获得了不错的推广应用,包括:C4.5决策树模型[5]、模糊系统结合遗传算法(Genetic Algorithm, GA)模型[6]、人工神经网络(Artificial Neural Network, ANN)结合诊断[7]、矢量量化网络(Learning Vector Quantization, LVQ)[8]、贝叶斯分类[9]、支持向量机(Support Vector Machine, SVM)[10]诊断等。其中:Quinlan[5]采用十倍交叉验证的C4.5决策树方法获得了94.74%的分类识别率;Hamilton等[11]利用近似分类的规则归纳(Rule Induction through Approximate Classification, RIAC)方法进行乳腺癌检测获得94.94%的正确率;Abonyi等[12]以监督聚类技术对乳腺癌数据处理后获得的识别率是95.57%;Karabatak等[13]将ANN技术与关联规则结合,最终的分类准确率为95.6%;而Sewak等[10]通过基于SVM的乳腺癌分类方法可以获得98.56%识别率。相对于其他机器学习方法,SVM因其优越的泛化能力,在乳腺癌的疾病诊断中获得良好应用。因此,本文主要研究SVM在乳腺癌的疾病诊断中的应用。

收稿日期:2015-06-01;**修回日期**:2015-07-05。

基金项目:国家自然科学基金资助项目(61373093);江苏省自然科学基金资助项目(BK20140008,BK201222725);江苏省高校自然科学研究项目(13KJA520001);江苏省"青蓝工程"资助项目;苏州大学第17届大学生课外学术科研基金资助项目(KY2015545B)。

作者简介:卢星凝(1992-),女,江苏淮安人,硕士研究生,主要研究方向:机器学习、模式识别;　张莉(1975-),女,江苏张家港人,教授,博士生导师,CCF高级会员,主要研究方向:机器学习、模式识别、图像处理。

由于乳腺癌数据属性存在冗余性,进行属性约简有利于提升诊断效果。因此,Chen 等[3]提出将属性约简与支持向量机(Attribute Reduction and SVM, AR-SVM)结合应用到乳腺癌的诊断中[3]。利用粗糙集理论(Rough Set, RS)对乳腺癌特征属性进行去冗余操作,通过最小化属性个数和区分矩阵进行属性约简。但是该方法没有考虑到整体的癌症条件属性对决策属性的依赖度,可能造成弱相关或不相关属性的存在,降低诊断正确率。

针对约简的特征属性可能产生不相关或冗余的问题,本文提出一种基于属性约简和支持向量机集成的乳腺癌诊断决策方案——MAR-SVM(Modified AR-SVM)。该方案中,采用遗传算法[14]对乳腺癌属性进行约简,对约简后的多特征子集用支持向量机集成实现诊断。在遗传算法中,本文设计了新的适应度函数。该适应度函数结合了最小化属性个数、最大化区分矩阵可区别属性的个数以及约简属性对决策属性的依赖度。在遗传算法迭代到一定次数后,种群中存在较优的多个可行解,获得多个约简属性子集;每个约简属性子集用一个SVM进行诊断训练;最终的诊断结果是多个SVM分类器的集成结果。本文方法考虑了条件属性对决策属性的依赖度,尽可能排除不相关的冗余属性,同时利用分类器集成有效提高对乳腺癌数据诊断决策的效率。

1 相关研究

1.1 粗糙集

粗糙集一般是对信息表或信息系统形式的数据进行处理[15]。信息表 S 一般可以表达为 $S=(U,A,V,f)$。其中:$U=\{\boldsymbol{x}_i\}_{i=1}^{n}$,称为论域;$A=\{a_j\}_{j=1}^{m}$ 是属性 a_j 的有限集合,且 $A=B\cup D$,$B\cap D=\varnothing$,B 表示条件属性集合,D 是决策属性集合;$V=\{V_j\}_{j=1}^{m}$,V_j 表示属性 a_j 的值域;$g:U\times A\rightarrow V$ 表示信息函数,满足 $g(\boldsymbol{x}_i,a_j)\in V_j$。

令信息系统 S 中 $|U|=n$,则 S 的区分矩阵是一个 $n\times n$ 的矩阵,满足 $a(\boldsymbol{x},\boldsymbol{x}')=\{a\in A\mid g(\boldsymbol{x},a)\neq g(\boldsymbol{x}',a)\}$,$\boldsymbol{x}$ 和 $\boldsymbol{x}'$ 分别表示 U 中不相同的两个对象,$a(\boldsymbol{x},\boldsymbol{x}')$ 是指区别对象 $\boldsymbol{x}$ 和 $\boldsymbol{x}'$ 的所有属性的集合。

RS中正域的概念来源于近似的划分。所谓近似,是通过一个已有集合 U 的一个划分中的集合的并集对 U 的一个任意子集 X 进行逼近,作为 X 的近似。设集合 $X\subseteq U$,R 为论域 U 上的等价关系,X 的下近似集 $\underline{apr}(X)$ 表示 U 中的对象一定在 X 中,即

$$\underline{apr}(X)=\{\boldsymbol{x}\in U\mid[\boldsymbol{x}]_R\subseteq X\} \tag{1}$$

其中:$[\boldsymbol{x}]_R$ 表示对象 $\boldsymbol{x}$ 在等价关系 R 下的等价类。正域 $POS(X)$ 中的元素一定属于 X,其表达式为:

$$POS(X)=\underline{apr}(X) \tag{2}$$

1.2 基于SVM的分类

1995年,Cortes 和 Vapnik 首先提出了支持向量机[16]。在统计学的VC维(Vapnik-Chervonenkis Dimension)理论以及结构风险最小概念基础上,SVM将向量映射到高维空间,并在该空间建立分隔超平面将数据分开,使得这个超平面具有最大的边缘。

在分类问题中,假设有训练数据集 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$,其中:$\boldsymbol{x}_i\in\mathbf{R}^m$ 表示 m 维数据特征,$y_i\in\{-1,+1\}$ 代表数据类别,n 是数据个数。SVM通过最小化下面的优化问题来解决两类分类问题:

$$\min_{w,b,\xi}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.}\quad y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)\geqslant 1-\xi_i;\ \xi_i\geqslant 0,\ i=1,2,\cdots,n \tag{3}$$

其中:$\boldsymbol{w}\in\mathbf{R}^m$ 是超平面的法向量,b 是阈值,ξ_i 是引入的松弛变量,$C>0$ 是惩罚因子。采用拉格朗日乘子方法和核技巧,求得优化问题(6)的对偶规划:

$$\max\quad\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{4}$$

$$\text{s.t.}\quad\sum_{i=1}^{n}\alpha_i y_i=0;\ 0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,n$$

其中:α_i 是拉格朗日乘子,$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 是满足 Mercer 条件的核函数。SVM的决策函数可以表示为:

$$f(\boldsymbol{x})=\operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{5}$$

其中:sgn 表示二值函数,其值为 $\{-1,+1\}$。

2 MAR-SVM模型

MAR-SVM和AR-SVM都是利用遗传算法对乳腺癌数据进行属性约简,得到约简属性集后再用SVM进行训练处理。和AR-SVM相比,MAR-SVM主要有两个特点:一是属性约简中适应度函数的设计,二是分类器采用集成的方式。

2.1 属性约简

遗传算法的算法流程如图1所示。在遗传算法属性约简中,采用二进制串编码表示长度为 m 的属性编码,即每一条编码数据的属性长度为 m。在二进制编码中,用0表示该属性被约简而不考虑,1表示该属性需要被保存。简要的算法描述如下:首先随机生成初始种群 $\{r_i\}_{i=1}^{G}$ 和定义一个适应度函数 $h(r)$,其中 $r_i\in\{0,1\}^m$ 表示一个个体,包含 m 位二进制,G 是种群的大小。其次根据适应度函数计算种群中每一个体的适应度。适应度的大小表明了个体的优劣,值越大则个体越可能是最优解。若满足终止条件,则算法终止;否则进行选择、交叉和变异操作,重复产生新的种群并计算适应度函数。选择算子采用轮盘赌的方式,且将种群中适应度最佳值的个体替代其中最差值的个体;交叉运算是以一定的概率选择个体参与交叉,变异运算是以一定概率选择个体进行变异[14]。

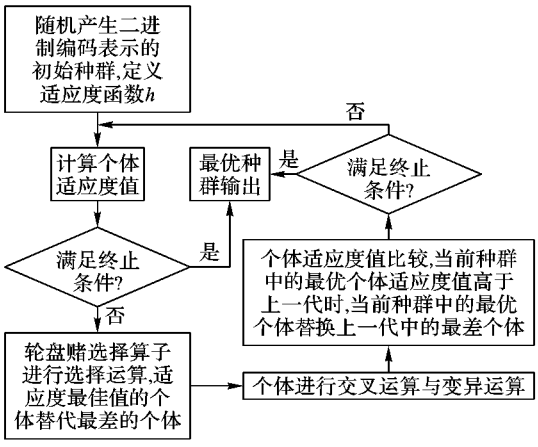

图1 遗传算法框架

通常情况下,对属性约简的过程往往是去除冗余属性的过程,为了对属性进行尽可能简化去冗余,前面提到AR-SVM中将最小化属性个数与区别矩阵约简方法结合。该方法和

MAR-SVM 都采用了遗传算法来进行属性约简,不同之处在于适应度函数的设计。AR-SVM 提出的基于粗糙集约简方法中,通过最小化属性个数和最大化区分矩阵可区别属性的个数作为适应度函数进行属性约简,其适应度函数表示如下:

$$h(r) = \frac{m - m_r}{m} + \frac{O_r}{M} \tag{6}$$

其中:m 是该条属性编码的总长度,m_r 是编码中标记为 1 的属性的个数;O_r 为保留的属性集与区分矩阵中元素进行合取后不为 0 的元素个数,表示该属性编码能区分的对象数量;M 为区分矩阵中的元素个数。

在适应度函数的设计中,除了考虑最小化属性个数和区分矩阵约简之外,还考虑了约简属性集与决策属性之间的依赖度。约简属性对决策属性的依赖度也是一种属性约简方法[17],为了提高约简的效率,本文在式(6)中引入了约简属性对决策属性的依赖度。在新方法中,适应度函数定义为:

$$h(r) = \frac{Pos(B_r, D)}{n} + \frac{m - m_r}{m} + \frac{O_r}{M} \tag{7}$$

其中:B_r 表示数据根据个体 r 约简而得到的新的属性数据集,D 表示决策属性,n 表示数据的总数目。适应度函数值 $h(r)$ 越大,个体 r 成为最优解的可能性也就越大。

满足了终止条件后,会得到包含最优解的种群 $\{r_i\}_{i=1}^{G}$。这里按照个体的适应度值进行从大到小的排序,不妨设 $h(r_1) \geqslant h(r_2) \geqslant \cdots \geqslant h(r_G)$。取前 N 条个体组成 N 个条件属性子集,$B_j = \{a_i \mid r_j^i = 1, a_i \in A\}$, $j = 1,2,\cdots,N$,其中 r_j^i 表示第 j 个个体的第 i 位。通常,令 N 是一个奇数[18]。

2.2　诊断决策模型

基于属性约简和支持向量机集成的诊断决策模型如图 2 所示,包括了 3.1 节的遗传算法属性约简部分。在利用遗传算法获得了 N 组条件属性子集后,采用集成多个 SVM 的方式来进行诊断测试。对条件属性子集 B_j,本文采用一个 SVM 子分类器(5)来训练,得到其判别模型 f_j,即

$$f_j(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i^j y_i K(\boldsymbol{x}_i^j, \boldsymbol{x}^j) + b^j\right); \ j = 1,2,\cdots,N \tag{8}$$

其中:α_i^j 和 b^j 是 SVM 判别模型 f_j 的参数,$\boldsymbol{x}_i^j$ 和 $\boldsymbol{x}^j$ 分别表示的是 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}$ 只取条件属性子集 B_j 中属性。对于一个未见样本 $\boldsymbol{x}$,需要计算它在这 N 个判别模型中的判别值,即得到 $f_j(\boldsymbol{x})$。最终的决策结果取决于下面的表达式:

$$\hat{y} = \operatorname{sgn}\left(\sum_{j=1}^{N} f_j(\boldsymbol{x})\right) \tag{9}$$

其中:$\hat{y}$ 是 $\boldsymbol{x}$ 的诊断结果。

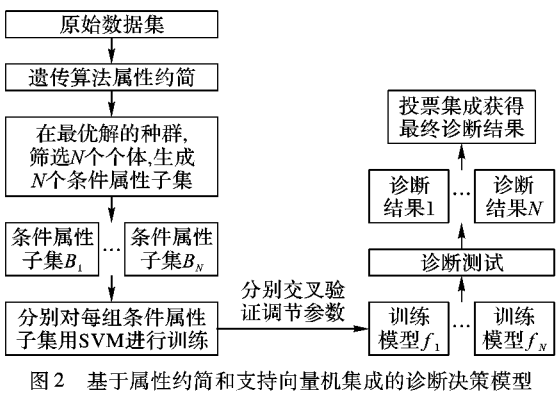

图 2　基于属性约简和支持向量机集成的诊断决策模型

3　实验数据分析

本章通过在 Wisconsin 癌症数据集(Wisconsin Breast Cancer Database, WBCD)的训练与测试验证本文算法的效果。

3.1　数据库和数据处理

WBCD 数据库是从人类乳腺癌组织中细针穿刺搜集得到的[19],共包括 683 个样本病例,其中 444 个样本被医学专家归为良性,239 个样本被归为恶性。每个样本病例分别由 9 个特征属性 $\{a_1, a_2, \cdots, a_9\}$ 组成,每个属性皆由 1 到 10 之间的整数表示。样本属性如下所示:

a_1:肿块密度(clump thickness);
a_2:细胞大小均匀性(uniformity of cell size);
a_3:细胞形状均匀性(uniformity of cell shape);
a_4:边界粘连(marginal adhesion);
a_5:单个上皮细胞大小(single epithelial cell size);
a_6:裸核(bare nucleoi);
a_7:微受激染色质(bland chromatin);
a_8:正常核(normal nuclei);
a_9:有丝分裂(mitoses)。

本文对乳腺癌数据进行了归一化预处理,把所有的数据归一化到区间[-1,1],以缩小属性值差异,优化实验效果。对属性 a_j,归一化公式如下:

$$x_j = 2\left(\frac{x_j - a_{j,\min}}{a_{j,\max} - a_{j,\min}}\right) - 1 \tag{10}$$

其中,x_j 是样本 x 在属性 a_j 下的取值,$a_{j,\max}$ 代表属性 a_j 的所有取值的最大值,$a_{j,\min}$ 代表属性 a_j 的所有取值的最小值。

3.2　训练集和测试集

生成训练集和测试集的过程中,考虑到乳腺癌数据集中良性和恶性数据的分布比例,采用分层抽样法的方法,分别对训练集与测试集按照 50%∶50%、70%∶30% 和 80%∶20% 三种比例进行划分。表 1 是乳腺癌数据集的训练集与测试集的比例分配情况。

表 1　训练集与测试集分配情况

训练集∶测试集比例	训练集样本数目	测试集样本数目
50%∶50%	342	341
70%∶30%	478	205
80%∶20%	546	137

3.3　方法设置和性能评估

遗传算法采用的 Matlab 代码编程实现,SVM 的训练和测试利用林智仁(Lin Chih-Jen)教授公布的 LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools)得以实现。SVM 的核参数以及正则参数采用交叉验证的方法来调节。设定高斯核参数的取值范围为 $\{2^{-15}, 2^{-13}, \cdots, 2^{1}\}$,正则因子的取值范围为 $\{2^{-5}, 2^{-3}, \cdots, 2^{15}\}$,通过在训练集上交叉验证获得最优参数,生成判别模型。

早期诊断措施准确性的性能评价包括敏感性和特异性两大指标。对诊断决策模型识别率的比较过程中,本文进行敏感性(sen)、特异性(spe)和准确率($accuracy$)的计算,公式如下:

$$sen = \frac{TP}{TP + FN} \tag{11}$$

$$spe = \frac{TN}{TN + FP} \tag{12}$$

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

其中:*TP* 表示真阳性,即正确被分类的恶性样本的数目;*FP* 表示假阳性,即被错分类为恶性的良性样本数目;*TN* 表示真阴性,即被正确分类的良性样本的数目;*FN* 表示假阴性,即被错分类为良性的恶性样本数目。除此之外,本文还用了混淆矩阵来对比方法的性能,表2是数据诊断决策的混淆矩阵的表示。

表2 混淆矩阵

数据集结果(良性)	数据集结果(恶性)	实验决策结果
TN	*FN*	良性
FP	*TP*	恶性

3.4 实验结果

实验中对比了三种方法:不做属性约简的SVM方法、AR-SVM,以及本文提出的MAR-SVM。

在80%:20%比例下,MAR-SVM挑选出来的特征如表3所示。和文献[18]一样,令条件属性子集个数 $N = 5$。在这5个条件属性子集中,都没有包括 a_1 和 a_2,也就是说肿块密度和细胞大小均匀性被MAR-SVM方法约简了;而其他的7个特征分别以不同的组合形式出现在子集中。AR-SVM所挑选出来的属性是由文献[3]所提供的。

表3 MAR-SVM以及AR-SVM生成的约简属性子集

方 法	属性子集
MAR-SVM	$B_1 = \{a_3, a_5, a_6\}$ $B_2 = \{a_3, a_4, a_8\}$ $B_3 = \{a_3, a_4, a_7, a_8\}$ $B_4 = \{a_4, a_6, a_8, a_9\}$ $B_5 = \{a_6, a_7, a_8\}$
AR-SVM[3]	$\{a_1, a_3, a_4, a_6, a_9\}$

随后的实验按照表3给出的属性集,分别用SVM来学习。依次按照三种比例随机分割数据(训练集:测试集分别为80%:20%、70%:30%和50%:50%),本文给出三种不同方法的混淆矩阵,分别对应了真阳性、假阳性、真阴性和假阴性四种情况下数据的分布。表4~6对应于不同的数据分配比例。在80%:20%情况下(表4),三种方法的 *TN* 的个数均为88,但是MAR-SVM的 *TP* 个数最多,这样就会提高MAR-SVM的诊断性能。相似地情形出现在70%:30%的比例中(表5)。在50%:50%情况下(表6),三种方法的 *TP* 的个数差不多,但是MAR-SVM的 *TN* 个数最多,这样也会提高MAR-SVM的诊断性能。而为了进一步地对文中基于GA属性约简的SVM方法作出客观评价,本文增加了在训练集与测试集50%:50%下主成分分析(Principal Component Analysis, PCA)方法实验性能,表7是PCA约简后经过SVM在训练集与测试集在50%:50%下的混淆矩阵。可以看出,与PCA-SVM方法相比较,MAR-SVM在 *TP* 个数上有领先优势,即敏感性表现较好;而 *TN* 个数上有不足,即特异性表现有待提高;但是总体识别率上,MAR-SVM仍然优于PCA-SVM方法。

本文随机分割数据10次,给出10次的平均结果。表8~10分别给出了基于GA属性约简和未进行属性约简这三种方法在三种比例下的诊断性能。通过比较发现,本文提出的决策诊断系统在诊断准确率上有明显优势,而敏感性和特异性也有较好的性能,尤其在敏感性中有不错的表现。训练集:测试集在80%:20%(表8)的情况下,由于其在数据集上的充分训练,整体分类识别率明显优于其他划分比例。综合三种情况下,本文提出的决策模型相比其他方法而言准确率至少高出约2%,甚至5%。在特异性指标下,经过属性约简后的性能要优于没有属性约简的,在敏感性和准确率指标下,AR-SVM不一定优于SVM,也就是说单一的属性约简不一定有效。MAR-SVM是采用多种属性约简方案,最后把这些方案所能带来优势进行了融合,因而本文方法要优于SVM。

表4 训练:测试集为80%:20%下的混淆矩阵

方法	数据集结果		实验决策结果
	良性	恶性	
MAR-SVM	88	1	良性
	1	47	恶性
AR-SVM	88	4	良性
	1	44	恶性
SVM	88	2	良性
	1	46	恶性

表5 训练:测试集为70%:30%下的混淆矩阵

方法	数据集结果		实验决策结果
	良性	恶性	
MAR-SVM	129	4	良性
	2	70	恶性
AR-SVM	128	5	良性
	5	67	恶性
SVM	129	5	良性
	4	67	恶性

表6 训练:测试集为50%:50%下的混淆矩阵

方法	数据集结果		实验决策结果
	良性	恶性	
MAR-SVM	214	4	良性
	8	115	恶性
AR-SVM	207	3	良性
	15	116	恶性
SVM	212	4	212
	10	115	10

表7 MAR-SVM在50%:50%比例下PCA混淆矩阵

方法	数据集结果		实验决策结果
	良性	恶性	
MAR-SVM	218	11	良性
	4	108	恶性

表8 三种方法在80%:20%数据集比例下的性能 %

方法	敏感性	特异性	准确率
MAR_SVM	89.59	98.06	95.09
AR_SVM	78.78	99.01	92.92
SVM	79.15	98.62	91.80

表9 三种方法在70%:30%数据集比例下的性能 %

方法	敏感性	特异性	准确率
MAR_SVM	83.91	97.00	92.41
AR_SVM	76.87	96.93	89.88
SVM	79.35	96.42	90.43

计算机应用 www.joca.cn

表 10　三种方法在 50%:50% 数据集比例下的性能　%

方法	敏感性	特异性	准确率
MAR_SVM	75.74	97.28	89.76
AR_SVM	61.74	97.26	84.86
SVM	68.85	96.61	86.92

4　结语

通过在乳腺癌数据集上不同划分数据集的多个实验,有效地证明了基于属性约简和 SVM 集成的方法对乳腺癌数据集的准确识别。在成功约简冗余数据集之后,明显地提高了乳腺癌疾病的识别准确率、敏感性和特异性。

然而在仿真过程中发现,尽管本文提出的决策模型在诊断过程中总体效率有显著提高,但是在敏感性和特异性的识别率略有差距,如何在实现高识别率、高特异性同时,更进一步提高诊断结果的敏感性,还需在下一步的研究中继续探索。

参考文献:

[1] SIEGEL R, MA J, ZOU Z H, JEMAL A. Cancer statistics, 2014 [J]. A Cancer Journal for Clinicians, 2014, 64(1): 9 - 29.

[2] WEST D, MANGIAMELI P, RAMPAL R, *et al.* Ensemble strategies for a medical diagnosis decision support system: a breast cancer diagnosis application [J]. European Journal of Operational Research, 2005, 162(2): 532 - 551.

[3] CHEN H L, YANG B, LIU J, *et al.* A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis [J]. Expert System with Application, 2011, 38(7): 9014 - 9022.

[4] WOLBERG W H, MANGASARIAN O L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology [J]. Proceedings of the National Academy of Sciences, 1990, 87(23): 9193 - 9196.

[5] QUINLAN J R. Improved use of continuous attributes in C4.5[J]. Journal of Artificial Intelligence Research, 1996, 4(1): 77 - 90.

[6] PENA-REYES C A, SIPPER M. A fuzzy-genetic approach to breast cancer diagnosis [J]. Artificial Intelligence in Medicine, 1999, 17(2): 131 - 155.

[7] SETIONO R. Generating concise and accurate classification rules for breast cancer diagnosis [J]. Artificial Intelligence in Medicine, 2000, 18(3): 205 - 219.

[8] GOODMAN D E, BOGGESS L C, WATKINS A B. Artificial immune system classification of multiple-class problems [EB/OL]. [2014-10-10]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.4865.

[9] KONONENKO I. Inductive and Bayesian learning in medical diagnosis [J]. Applied Artificial Intelligence an International Journal, 1993, 7(4): 317 - 337.

[10] SEWAK M, VAIDYA P, CHAN C C, *et al.* SVM approach to breast cancer classification[C]// IMSCCS 2007: Proceedings of the Second International Multi-Symposiums on Computer and Computational Sciences. Piscataway: IEEE Press, 2007: 32 - 37.

[11] HAMILTON H J, SHAN N, CERCONE N. RIAC: a rule induction algorithm based on approximate classification [EB/OL]. [2014-10-10]. http://wiki.eecs.yorku.ca/course_archive/2014-15/F/4412/_media/9606.pdf.

[12] ABONYI J, SZEIFERT F. Supervised fuzzy clustering for the identification of fuzzy classifiers [J]. Pattern Recognition Letters, 2003, 24(14): 2195 - 2207.

[13] KARABATAK M, INCE M C. An expert system for detection of breast cancer based on association rules and neural network [J]. Expert Systems with Applications, 2009, 36(2): 3465 - 3469.

[14] WANG Y, TETKO I V, HALL M A, *et al.* Gene selection from microarray data for cancer classification — a machine learning approach [J]. Computational Biology and Chemistry, 2005, 29(1): 37 - 46.

[15] YAO Y. Decision-theoretic rough set models [C]// RSKT 2007: Proceedings of the Second International Conference on Rough Sets and Knowledge Technology, LNCS 4481. Berlin: Springer-Verlag, 2007: 1 - 12.

[16] NOBLE W S. What is a support vector machine [J]. Nature Biotechnology, 2006, 24(12): 1565 - 1567.

[17] YAMAGUCHI D. Attribute dependency functions considering data efficiency [J]. International Journal of Approximate Reasoning, 2009, 51(1): 89 - 98.

[18] POLAT K, GUNES S. Breast cancer diagnosis using least square support vector machine [J]. Digital Signal Processing, 2007, 17(4): 694 - 701.

[19] SETIONO R. Generating concise and accurate classification rules for breast cancer diagnosis [J]. Artificial Intelligence in medicine, 2000, 18(3): 205 - 219.

(上接第 2780 页)

[31] XIA W, WANG L. Research on and implementation of parallel ant colony algorithm based on MapReduce[J]. Electronic Science and Technology, 2013, 26(2): 146 - 149. (夏卫雷, 王立松. 基于 MapReduce 的并行蚁群算法研究与实现[J]. 电子科技, 2013, 26(2): 146 - 149.)

[32] BIAN H, CHEN Y, DU X, *et al.* Equivalent connection optimization based on Spark[J]. Journal of East China Normal University: Natural Sciences, 2014(5): 263 - 270. (卞昊穹, 陈跃国, 杜小勇, 等. Spark 上的等值连接优化[J]. 华东师范大学学报: 自然科学版, 2014(5): 263 - 270.)

[33] QIU R. Implement and application of CURE algorithm based on Spark[D]. Guangzhou: South China University of Technology, 2014. (邱荣财. 基于 Spark 平台的 CURE 算法并行化设计与应用[D]. 广州: 华南理工大学, 2014.)

[34] TANG Z. The design and implement of a machine learning platform based on Spark[D]. Xiamen: Xiamen University, 2014. (唐振坤. 基于 Spark 的机器学习平台设计与实现[D]. 厦门: 厦门大学, 2014.)

[35] STUTZLE T, HOOS H. MAX-MIN ant system and local search for the traveling salesman problem[C]// ICEC 1997: Proceedings of the 1997 IEEE International Conference on Evolutionary Computation. Piscataway: IEEE Press, 1997: 309 - 314.

[36] REINELT G. TSPLIB — a traveling salesman problem library[J]. ORSA Journal on Computing, 1991, 3(4): 376 - 384.

[37] XING H, QU R, KENDALL G, *et al.* A path-oriented encoding evolutionary algorithm for network coding resource minimization [J]. Journal of the Operational Research Society, 2013, 65(8): 1261 - 1277.

[38] CUI W, LI X, ZHOU S, *et al.* Investigation on process parameters of electrospinning system through orthogonal experimental design [J]. Journal of Applied Polymer Science, 2007, 103(5): 3105 - 3112.

计算机应用 www.joca.cn