文章编号：2095-2163(2020)11-0097-05 中图分类号：TP181 文献标志码：A

基于系统聚类和 SVM 模型的乳腺癌诊断研究

余 莹[1]，樊重俊[1]，朱人杰[1,2]，熊红林[1,3]
（1 上海理工大学 管理学院，上海 200093；2 同济大学附属东方医院，上海 200120；3 万达信息股份有限公司，上海 201112）

摘 要：随着医疗技术的发展，临床医学中已收集了用于乳腺癌诊断的不同肿瘤特征。然而如何从庞大的医疗数据集中选择特征信息，以支持临床疾病诊断，是一项艰巨而耗时的任务。针对于此，本文提出了基于系统聚类和支持向量机(H-SVM)的组合模型。其中系统聚类算法用于特征选择，分别识别良性肿瘤和恶性肿瘤的隐藏模式；通过从属函数计算原始肿瘤数据与隐藏模式之间的相似度进行特征重建；重建后的数据集作为新的特征集通过支持向量机算法训练分类器，以检验分类效果。实验结果表明，该算法从威斯康星州乳腺癌(WDBC)数据集训练阶段的 32 个原始特征中提取了 15 个抽象的肿瘤特征，不仅将分类精确率提高到 97.50%，而且大大减少了模型训练时间。

关键词：系统聚类；特征选择；支持向量机；乳腺癌诊断

Breast Cancer Diagnosis Based on Combined H-SVM Model

YU Ying[1], FAN Chongjun[1], Zhu Renjie[1,2], XIONG Honglin[1,3]
(1 Business School, University of Shanghai for Science and Technology, Shanghai 200093, China;
2 Shanghai East Hospital Affiliated to Tongji University, Shanghai 200120, China;
3 Wonders Information Co., Ltd., Shanghai 201112, China)

【Abstract】 With the development of medical technology, different tumor characteristics for the diagnosis of breast cancer have been collected in clinical medicine. However, how to select feature information from a huge medical data set to support clinical disease diagnosis is a difficult and time-consuming task. In order to extract useful information in the data set and diagnose breast cancer, a combined model based on hierarchical clustering and support vector machine (H-SVM) is proposed, in which the hierarchical clustering algorithm is used for feature selection , to identify hidden patterns of benign and malignant tumors respectively, and then calculate the similarity between the original tumor data and the hidden pattern through the membership function to perform feature reconstruction. The reconstructed data set is used as a new feature set to train the classifier through the support vector machine algorithm to test the classification effect. The results show that the algorithm extracts 15 abstract tumor features from the 32 original features in the training stage on the Wisconsin Breast Cancer (WDBC) dataset, which not only improves the classification accuracy to 97.50%, but also greatly reduces the model training time .

【Key words】 Hierarchical clustering; Feature selection; SVM; Breast cancer diagnosis

0 引 言

近年来乳腺癌的多发以及所带来的严重后果已经在全球范围内引起了广泛关注，乳腺癌是影响成年女性的主要慢性疾病之一。全球范围内每年都有约 1 000 万的女性被诊断出罹患乳腺癌，并且超过 50 万女性死于乳腺癌[1]。随着现代经济的发展和医疗技术的进步，有大量的资源和现代技术可以应用于乳腺癌的筛查、诊断和控制工作。对于医生来说，要从大量的癌症病例当中详细了解每一个癌症患者的特征是十分困难的。因此，数据分析方法可以成为医生做出癌症诊断决策时的重要助手[2]。

早在 1999 年，Pena-Reyes 和 Sipper[3] 提出了一种模糊遗传算法诊断乳腺癌。其研究结果表明，数据挖掘技术已成功应用于癌症预测中，传统的乳腺癌诊断已转化为数据分析领域的分类问题。现有的乳腺癌数据集被分为良性和恶性两类，通过历史肿瘤数据训练得到合适的分类器，来预测新的肿瘤数据。但随着描述肿瘤特征数据的增加，分类器的计算时间也急剧增加，在这种情况下，乳腺癌诊断的基本要求不仅是准确性，还包括时间复杂度。考虑到时间效率，如何从庞大的数据集中挖掘和提取必要的信息、过滤特征成为一个新的问题。

Akay(2009)[4] 提出了一种基于 SVM 与特征选择相结合的方法来进行乳腺癌诊断。通过使用 F 分数[5] 来计算特征价值，选择原始肿瘤特征的最佳

作者简介：余 莹(1995-)，女，硕士研究生，主要研究方向：大数据分析；樊重俊(1963-)，男，博士，教授，博士生导师，主要研究方向：大数据分析。
通讯作者：樊重俊 Email：Fan.chongjun@163.com

收稿日期：2020-08-16

子集进行 SVM 训练。

Akay(2009)[4]提出了一种基于 SVM 与特征选择相结合的方法来进行乳腺癌诊断,通过使用 F 分数[5]来计算特征价值。进而为了找到最佳的参数设置组合,使诊断准确率达到最高,进行了耗时较长的网格搜索,选择原始肿瘤特征的最佳子集进行 SVM 训练。Prasad、Biswas 和 Jain(2010)[6]尝试了启发式算法和 SVM 的组合,以找出用于 SVM 训练的最佳特征子集。但是,这些方法的共同缺陷是,仅仅使用分类精确率作为评估不同特征选择方法的标准,而忽视了对不同子集进行详尽训练,以获得具有最佳诊断精确率的最优子集所消耗的大量模型训练时间。

因此,本文提出了基于系统聚类和支持向量机的组合模型。系统聚类算法作为一种无监督学习算法提取肿瘤特征,以识别肿瘤数据的隐藏模式,只在原始特征空间上进行聚类,不仅可以以更加紧凑的方式保留所有单个特征信息,而且避免了在不同子集上进行迭代训练,以节约模型训练时间。基于特征选择的结果,应用从属函数计算这些隐藏模式与每个肿瘤之间的相似性,并将其作为新的特征对原始肿瘤数据进行特征重建,最后应用 SVM 算法对重建后的数据集进行分类。

1 研究方法

1.1 基于系统聚类的特征选择方法

系统聚类,也称层次聚类,是统计学方法中的一种聚类算法,其原理简单。首先,将所有样本本身归为一类,类与类之间的距离就是它们所包含的样本之间的距离;然后找出距离最近的两个类将它们合并为一个类,重新计算新生成的类与旧类之间的距离;不断重复以上步骤直到所有样本归为一类[7]。本文采用欧式距离计算距离矩阵,并采用离差平方和法判断类与类之间的距离。基于方差分析的思想是:如果分类正确,则分类结果应该满足,同类样本之间离差平方和较小,而异类样本之间离差平方和较大。

特征选择过程也可描述为数据转换过程,是将特征数据转化为定量的数据结构,以方便训练模型的过程。特征选择在具有高维特征空间的大规模数据中起着重要的作用。当训练数据为高维数据时,这个过程可以用来消除不必要的训练信息,在保持训练精度的同时,缩短总体训练时间[8]。特征选择的原则是,在不影响后续分类分布结果,不降低准确率及提取的特征子集应为稳定且适应度强的集合基础上,提取尽可能小的特征子集。在统计学中,特征选择的统计模型一般使用数学统计模型建立,以数学方程式的形式表示变量之间的函数关系。通过计算模型的残差平方和大小,评价模型的拟合程度。在对原始数据进行系统聚类后,需要对聚类结果进行相似性度量,从而决定最佳类的个数,相似性度量的方法如式(1)、式(2)[9]所示:

$$d_{avg}=\frac{\sum_{k=1}^{K}\sum_{i\in s_k}\sqrt{\sum_{j=1}^{F}(X_j^i-X_j^{\mu_k})^2}}{N}. \tag{1}$$

$$d_{\min}=\min\left\{\sqrt{\sum_{j=1}^{F}(X_j^{\mu_{k_1}}-X_j^{\mu_{k_2}})^2}\right\},\ \forall k_1\neq k_2. \tag{2}$$

其中,d_{avg} 是同一类 s_k 中每个成员 i 到质心 μ_k 的平均距离;$d_{\min}$ 表示任意两类质心之间的最小距离;X_j^i 表示成员 i 的第 j 个输入元素;$X_j^{\mu_k}$ 表示质心 μ_k 的第 j 个输入元素;N 是数据点的总数;F 是输入向量的维数。

最佳聚类数 K^*,通过使用如下方法求出最小有效率 θ 来获得,如式(3)所示[9]:

$$K^*=\arg\min_K\theta=\arg\min_K\frac{d_{avg}}{d_{\min}}. \tag{3}$$

其中,θ 是评估聚类数有效率的量值。θ 求得最小值的过程,也是每个成员与其簇质心的平均距离 d_{avg} 不断减小,而任意两个簇质心之间的最小距离 $d_{\min}$ 不断增加的过程。即在通过有效率 θ 求解最佳聚类数 K^* 的过程中,也满足了类内距离小、异类间距离大的条件。

当 K 的取值接近特征数目时,则无法找出隐藏模式;当 K 取值较小时,才会较明显地显示出隐藏模式。

1.2 特征重建

进行特征选择后,需在原始数据集的基础上进行特征重建。此时,未测试数据与之前步骤中选择出的新特征之间的相似程度,在新数据集的特征重建中扮演着重要的角色。因此,计算原始数据与各新特征之间相似性的从属函数极为重要。从属函数计算如式(4)、式(5)所示[9]:

$$f_c(X_j^i)=\begin{cases}1-\dfrac{|X_j^{\mu_c}-X_j^i|}{\max|X_j^{\mu_c}-X_j^n|},\\ \text{若 }\min(X_j^n)\leqslant X_j^i\leqslant\max(X_j^n),\ \forall n\in S_c;\\ 0,\ \text{其它}.\end{cases} \tag{4}$$

$$\rho_{ic} = \frac{1}{F}\sum_{j=1}^{F} f_c(X_j^i)\ ,\ 1 \leqslant c \leqslant K^m + K^b. \qquad (5)$$

其中，c 是新模式的指标，X_j^i 是原输入 i 的第 j 个特征，$X_j^{\mu_c}$ 是通过系统聚类得出的类 S_c 的中心 μ_c 的第 j 个特征，K^m 和 K^b 分别是通过系统聚类得出的良恶性隐藏模式的数目。

通过 ρ_{ic}，可刻画肿瘤 i 和肿瘤模式 S_c 之间的相似度程度，ρ_{ic} 的大小反映了二者的相似度，数值越大，相似度越高。将通过系统聚类提取的新模式作为肿瘤新的抽象特征，并通过从属函数计算所有原始肿瘤数据与肿瘤模式 S_c 之间相似程度，将其组成新数据，完成特征重建。

1.3　支持向量机分类

基于前两步的操作，数据的特征维度已经减小，并且具有新特征的数据集已经重建，可以应用传统的机器学习算法。由于支持向量机算法（SVM）自身的优势，对于线性可分的二分类问题，可通过找到一个最优分界面将两类分开；对于线性不可分的二分类问题，可利用核函数实现在高维特征空间分类。支持向量机算法在小样本、非线性及高维模式应用中具有优势，故本文选择支持向量机算法进行分类[11]：

$$\max_{\alpha}\left(\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j)\right),$$
$$\sum_{i=1}^{n}\alpha_i y_j = 0, \quad \forall\, 0 \leqslant \alpha_i \leqslant L. \qquad (6)$$

其中，x 是训练向量；y 是与训练向量相关的标签；α 是分类器超平面的参数向量；$K(\cdot)$ 为核函数；L 是由惩罚参数决定的错误分类数量。

2　实验及结果

2.1　乳腺癌数据描述

本文使用的数据来自加州大学尔湾分校的威斯康星州诊断性乳腺癌（WDBC）数据集。该数据集包含每个细胞核 10 个类别的 32 个特征，其分别是：半径、纹理值、周长、面积、光滑度、紧密度、凹度、凹点、对称性、分形维数。对于每个类别，分别测量 3 个指标：平均值、标准误差和最大值，包括样本的名称和类别一共 32 维，共包含 569 条数据，见表 1。

2.2　H-SVM 算法

使用 H-SVM 算法对乳腺癌数据进行诊断。为了对特征进行降维，分别在良性数据集和恶性数据集上使用特征选择方法提取肿瘤数据的隐藏模式，在判断最佳聚类数时，应用式（1）、（2）、（3）得到 K^*，在特征选择的基础上，利用式（4）、（5）进行特征重建，最后应用 SVM 算法进行分类。整个算法流程[10]如图 1 所示。

表 1　WDBC 数据集分布描述

Tab. 1　Summary of WDBC data attributes

类别	测量值		
	平均值	标准差	最大值
半径	6.98-28.11	0.112-2.873	7.93-39.04
纹理	9.71-39.28	0.36-4.89	12.02-49.54
周长	43.79-188.50	0.76-21.98	50.41-251.20
面积	143.50-2501.00	6.80-542.20	185.2-4254
平滑度	0.053-0.163	0.002-0.031	0.071-0.223
紧密度	0.019-0.345	0.002-0.135	0.027-1.058
凹度	0-0.427	0-0.396	0-1.252
凹点	0-0.201	0-0.053	0-0.291
对称性	0.106-0.304	0.008-0.079	0.157-0.664
分形维数	0.050-0.097	0.001-0.030	0.055-0.208

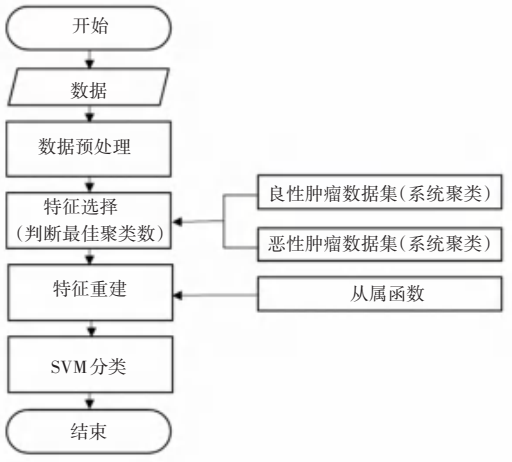


图 1　H-SVM 算法流程

Fig. 1　H-SVM algorithm flow

2.2.1　数据预处理

数据预处理过程主要包括二个方面，一是分离良性数据集与恶性数据集；二是数据标准化。

（1）良恶性数据集分离。由于在进行肿瘤隐藏模式识别时，良性肿瘤与恶性肿瘤的隐藏模式是分别存在的，而原数据集中良性肿瘤数据与恶性肿瘤数据则混合在一起。原数据中第二维为数据分类的标识，在进行数据集分离时只需按照 B（良性肿瘤数据集）或 M（恶性肿瘤数据集）筛选分离即可。

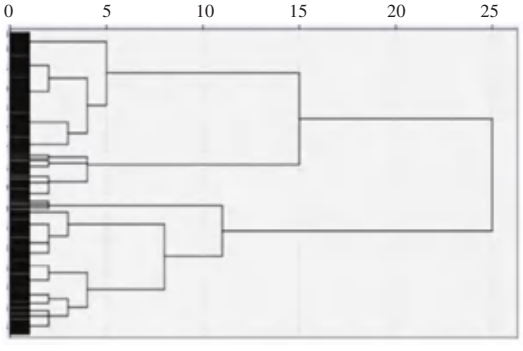
（2）数据标准化。在进行系统聚类分析前，需对数据集中标签属性进行归一化处理，以消除量纲对相似度的影响。即消除对聚类过程中相似矩阵计算的影响，从而获得一个更优的聚类结果。归一化公式如式（7）：

$$x_{ij}^{'} = \frac{x_{ij} - x_i^{\min}}{x_i^{\max} - x_i^{\min}}. \quad (7)$$

其中，i 为数据集的第 i 个属性；j 为数据集的第 j 条记录；x_{ij} 为数据集某属性原始记录；$x_i^{\max}$ 和 $x_i^{\min}$ 分别为数据集里第 i 个属性中的最大值和最小值。

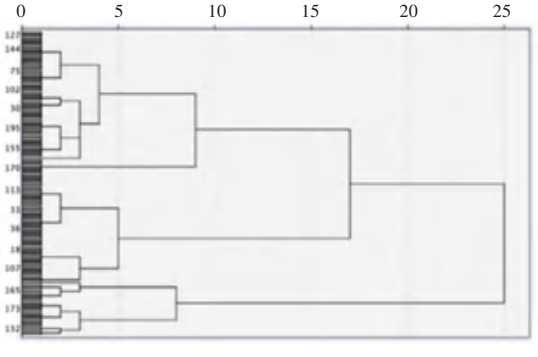
2.2.2 特征选择

首先，分别对良性肿瘤数据集与恶性肿瘤数据集进行系统聚类。图 2 为聚类结果谱系图（其中（a）为良性肿瘤数据聚类谱系图，（b）为恶性肿瘤聚类谱系图）。由图可见，系统聚类在良恶性肿瘤数据集上有很好的聚类效果，能够比较清晰地体现出类别的层次，即乳腺癌肿瘤数据的隐藏模式明显，各隐藏模式之间差距较大。



（a）良性肿瘤数据

（a）Benign tumor data



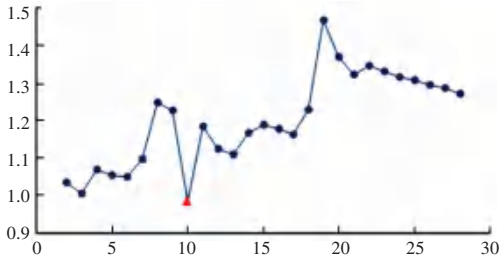
（b）恶性肿瘤数据

（b）Malignant tumor data

图 2 肿瘤数据系统聚类图

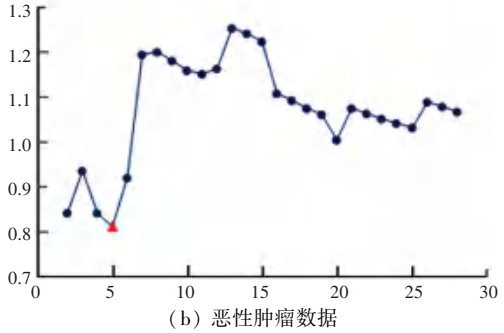
Fig. 2 Hierarchical graph

进行特征选择时，利用式（1）、（2）分别求得良恶性肿瘤数据对应的有效率，其中聚类数 K 的取值范围为（2,30）。聚类产生的每一类，代表一个肿瘤的隐藏模式；每一个类的类中心，代表该隐藏模式的类中心。利用式（3）求得每个簇的 θ 值，如图 3 所示。从图 3 中可以看出，在取值范围内，有效率 θ 有一个最小值。即当良性肿瘤类别数 $K^b = 10$ 时，θ^b 求得最小值；当恶性肿瘤类别数 $K^M = 5$ 时，θ^m 求得最小值。根据本文算法，以最紧凑的模式保留原始特征得到良、恶性肿瘤的最佳隐藏模式数分别为 10 种和 5 种。如图 5 所示。



（a）良性肿瘤数据

（a）Benign tumor data



（b）恶性肿瘤数据

（b）Malignant tumor data

图 3 肿瘤模式 K 值的确定

Fig. 3 Determine K for tumors

2.3 分类结果

分类算法结果的正确性用准确率来衡量，准确率越高说明分类的效果越好。本文 H-SVM 算法在 WDBC 数据集上应用的准确率为 96.5%。其计算公式为式（8）所示：

$$\text{accurary} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (8)$$

其中，TP 是真正数；TN 是真负数；FP 是假正数；FN 是假负数。

就准确率而言，本文提出的 H-SVM 算法与仅使用 SVM 算法进行分类比较，保证了高的预测精度；另一方面，H-SVM 算法是通过将原始数据进行特征选择以减少特征空间的维度，然后特征重建转换为新的数据集。从计算时间的角度来看，所提出的方法通过减少输入特征的数量，显著减少了训练时间。表 2 中将计算时间与传统的 SVM 算法进行了比较，显示了选择和提取特征的重要性。

表 2 结果比较

Tab. 2 Result comparison

	特征空间维度	精确率	训练时间
H-SVM	15	97.5%	0.208 8
SVM	30	95.3%	15.891 3[9]

（下转第 105 页）