doi:10.3969/j.issn.1672-5565.2015.02.10

高维蛋白质波谱癌症数据特征提取

吴文峰,刘毅慧*
(齐鲁工业大学信息学院,济南 250353)

摘　要:高维蛋白质波谱癌症数据分析,一直面临着高维数据的困扰。针对高维蛋白质波谱癌症数据在降维过程中的问题,提出基于小波分析技术和主成分分析技术的高维蛋白质波谱癌症数据特征提取的方法,并在特征提取之后,使用支持向量机进行分类。对8-7-02数据集进行2层小波分解时,分别使用db1、db3、db4、db6、db8、db10、haar小波基,并使用支持向量机进行分类,正确率分别达到98.18%、98.35%、98.04%、98.36%、97.89%、97.96%、98.20%。在进一步提高分类识别正确率的同时,提高了时间率。

关键词:小波分析;主成分分析;蛋白质波谱;降维;分类
中图分类号:Q629.73　　**文献标志码**:A　　**文章编号**:1672-5565(2015)-02-131-10

Feature selection for high-dimensional cancer protein mass spectrometry data

WU Wenfeng, LIU Yihui*
(*School of Information, Qilu University of Technology, Jinan* 250353, *China*)

Abstract: The analysis of high-dimensional cancer protein mass spectrometry data is full of trouble from high-dimensional data. We propose method for selecting the feature of high-dimensional cancer protein mass spectrometry data based on the wavelet analysis and principal component analysis, and solving the faled problems when we reduce the dimensionality of high-dimensional cancer protein mass spectrometry data. After feature selection, we use the support Vector Machine (SVM) for classification. We use wavelet decomposition on 8-7-02 data set at second level, use different wavelet basis (db1, db3, db4, db6, db8, db10, haar) and classify them with the SVM, then we get different recognition rates: 98.18%, 98.35%, 98.04%, 98.36%, 97.89%, 97.96%, 98.20%. Improve the classification accuracy and the efficiency of time simultaneously.

Keywords: Wavelet analysis; Principal component analysis; Protein mass spectrometry; Dimensionality reduction; Classify

近年来,蛋白质组学迅速发展。蛋白质波谱数据分析在癌症检测中得到了越来越广泛的应用。目前,在蛋白质波谱分析过程中,波谱信息主要通过基质辅助激光解析电离技术(Matrix-Assisted Laser Desorption/Ionization:MALDI)和表面增强激光解吸离子化飞行时间质谱技术(Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometry:SELDI-TOF-MS)来获得[1]。本文中的蛋白质波谱数据,主要是通过SELDI-TOF-MS技术得到。SELDI-TOF-MS技术主要由蛋白质芯片、飞行时间质谱仪和相关软件组成,其中蛋白质芯片是该技术的核心。

严勇等通过采用模式识别领域常用的决策树与AdaBoost技术来处理医学领域常用的质谱分析数据[2],研究了弱分类器个数对分类性能的影响,将AdaBoost与支持向量机进行类比。根据实验,从大间隔学习的角度,阐述了AdaBoost的优势。AdaBoost是二元分类方法中经常用到的一个提升方法[3]。AdaBoost对不同训练集训练时,采用同一个弱分类器,之后把在不同训练集上得到的分类器集合起来,组成一个更优的强分类器。邹修明等在基于蛋白质的癌症诊断实验中[4],通过基线校正和标

收稿日期:2015-01-14;**修回日期**:2015-03-10.
作者简介:吴文峰,男,硕士研究生,研究方向:智能信息及图像处理;E-mail:641178636@qq.com.
*通信作者:刘毅慧,女,博士,教授,研究方向:生物计算,智能信息处理等;E-mail:yxl@sdili.edu.cn.

准化,并使用分箱法对原始数据进行降维预处理,之后使用 T 检验方法来选取特征,对经过了一系列处理后的蛋白质质谱数据进行分析研究。论文中实验采用 10-fold 交叉验证和支持向量机对卵巢质谱数据集进行分类。杨合龙等针对如何有效分析高通量 SELDI-TOF 质谱数据以及筛选与肿瘤相关的蛋白质位点,提出一种基于近邻传播聚类分析 的特征选选择方法[5]。Kuehl B 等将蛋白质波谱分析应用于细菌生理研究,并结合主成分分析方法,区分细菌的不同生理状态[6]。EBERLIN L 等利用蛋白质波谱数据研究人类脑瘤,对正常和患病数据进行分类[7]。王昭鑫等针对癌症蛋白质谱数据中包含大量未知的内部结构和变量这一特点,在总结主元余像集主成分分析(二次主成分分析)应用的基础上,提出了应用 t-验证方法进行特征子集选取,之后用主元余像集主成分分析来提取特征,最后以线性判别分析进行分类的新方法[8]。

模式识别和分类的过程中,数据特征的质量对于识别和分类的速率和正确率至为重要。需要预先对数据进行降噪、降维、归一化等预处理,之后再提取特征,最后基于降维后的特征来进行模式的识别和分类。目前常用的数据降维降噪处理的方法有主成分分析法、T-test 法、Boosting、遗传算法、模拟退火算法、小波分析法等[9-14]。小波分析技术可以用于蛋白质质谱数据的分析,用它做降维去噪处理后的低频系数,可以有效的表征蛋白质质谱数据的特征信息。

高维数据的降维和特征提取方法研究依然很重要。本文将离散小波分析和主成分分析方法相结合,对几组癌症数据进行多维降噪处理,提取低频系数作为其特征数据。在小波分析过程中,对高维蛋白质波谱数据进行不同层的小波分解和基于不同小波基的分解,并做了详细的比较,选择出具有最佳识别率的分解层数和小波基。在之后的主成分分析过程中,通过实验比较,选择出最佳主成分。本文中将使用支持向量机对提取的特征数据进行分类。

1 相关理论

1.1 小波分析技术

时频分析,是时频联合域分析的简称。它提供了信号的时间域和频率域的联合信息,描述信号频率随着时间变化的关系。

小波分析是时频分析的一种,它在时域和频域里都能很好的表征局部信号特征,是一种多尺度信号分析方法。小波作为一重要的线性时频展开方法,不同于短时傅里叶(Fourier)变换,它是将信号展开为持续时间很短的高频基函数和持续时间较长的低频基函数,而这些不同的基函数是从单个原象小波通过平移和伸缩得到。小波又分两大类:连续小波和离散小波。

假设 $L2(R)$ 是 R 上平方可积函数所构成的函数空间。若 $\Psi(t) \in L2(R)$,并且其傅里叶变换 $\hat{\Psi}(\omega)$ 满足条件:

$$C_\Psi = \int_{-\infty}^{+\infty} = |\omega|^{-1} |\hat{\Psi}(\omega)|^2 \mathrm{d}\omega < \infty \quad 。 \quad (1)$$

称 Ψ 是一个基小波或者称作母小波,其中,R 为实数,t 为时间。把基小波伸缩和平移,可以得到一个小波序列

$$\Psi_{a,b}(t) = |a|^{-1/2} \Psi\left(\frac{x-b}{a}\right) \quad 。 \quad (2)$$

其中,a,b∈R,并且 a≠0。a 称为伸缩因子,b 称为平移因子。式子

$$(W_\Psi f)(a,b) = \langle f, \Psi_{a,b} \rangle = |a|^{-1/2} \int_{-\infty}^{+\infty} f(t) \Psi\left(\frac{t-b}{a}\right) \mathrm{d}t \quad 。 \quad (3)$$

定义为基小波 Ψ 的连续小波变换。$\bar{X}$ 为 X 的共轭运算。

在实际问题中,小波变换中的伸缩因子和平移因子往往都不是连续的,此时数值计算中需要采用离散小波变换。取 $a=a_0^m, b=nb_0a_0^m, m,n \in Z$ 代入式(2),得到相应的离散小波变换[15]:

$$(W_\Psi f)(a,b) = \langle f, \Psi_{a,b} \rangle = |a|^{-m/2} \int_{-\infty}^{+\infty} f(t) \Psi(a_0^{-m}t - nb_0) \mathrm{d}t \quad 。 \quad (4)$$

本文中,采用了离散小波变换,其中,Z 为整数。

小波分析中,选择一个小波基并确定一个小波分解的层次 N,然后对已知信号进行 N 层小波分解,如图 1 所示为小波分解示意图以及部分小波基,图 2、图 3 分别为小波分解前后数据信号波形。

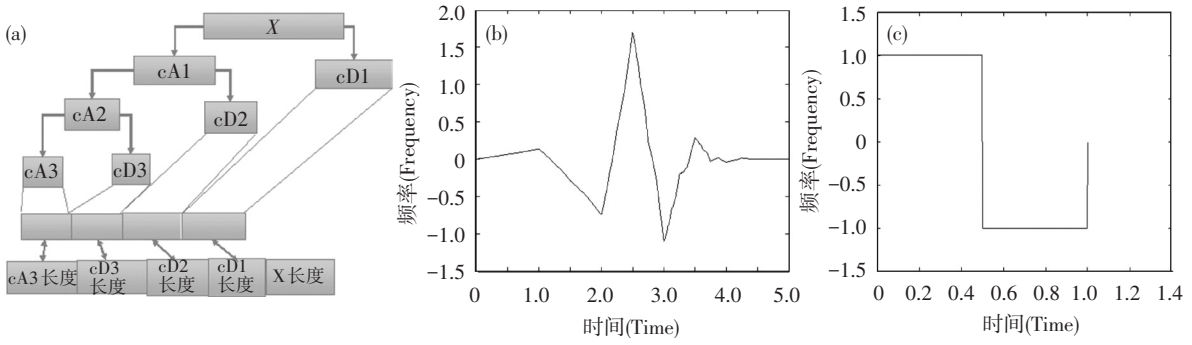

图 1　小波分解及小波基

Fig.1　Wavelet decomposition and wavelet basis

注：图(a)为小波分解：原始信号 X 经过一次分解后，得到高频系数 cD1 和低频系数 cA1。之后再次对低频系数进行分解，每次分解都会得到高、低频系数。图(b)为 db3 小波。图(c)为 haar 小波。

Notes：(a) is wavelet decomposition：fter first level wavelet decomposition on x，we get high frequency coefficient cD1 and low frequency coefficient cA1. Then decompose high frequency coefficient again，we will get high frequency coefficient and low frequency coefficient every decomposition. (b) is db3 wavelet basis. (c) is haar wavelet basis.

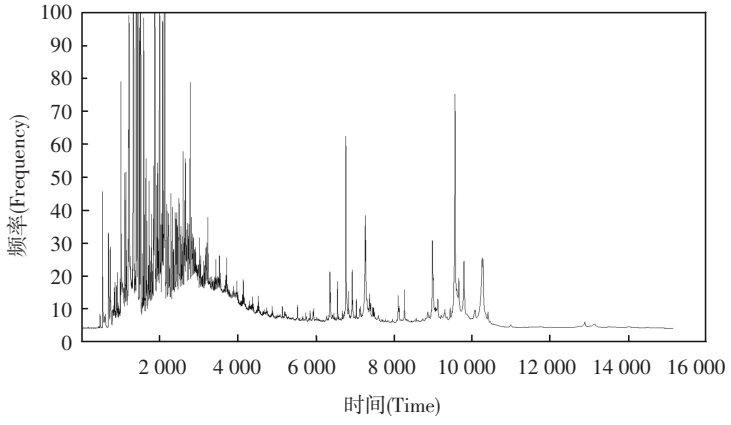

图 2　8-7-02 数据集第一组数据原始信号波形

Fig.2　The original waveform of the first series data of 8-7-02 data set

注：数据有 15 154 个属性，作为 Time 轴，属性值作为 Frequency 轴。

Notes：The original waveform of the first series data of 8-7-02 data set：the data set has 15 154 properties，set the properties as Time axis，property values as Frequency axis.

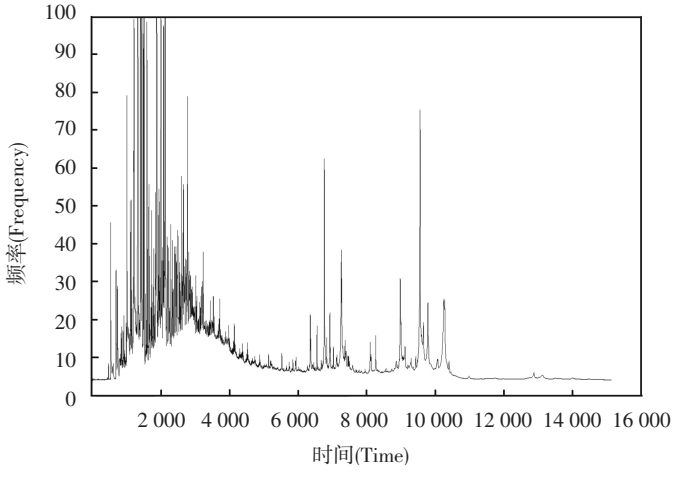

图 3　8-7-02 数据集第一组数据处理后波形

Fig.3　After fourth level wavelet decomposition on the first series data of 8-7-02 data set

注：采用 db3 小波基、4 层作为参数进行小波分解后波形，此时数据剩余 951 个属性。

Notes：After fourth level wavelet decomposition on the first series data of 8-7-02 data set，use db3 wavelet basis：the data has 951 properties now.

1.2 主成分分析

主成分分析(Principal component analysis, PCA)最早由皮尔逊(Pearson,1901)引入,后来由霍特林(Hotelling,1933)进一步发展。它是将多个线性相关变量压缩为少数几个不相关的变量的一种多元统计方法,最早由 Pearson 在研究对空间中的数据进行最佳直线和平面拟合时提出[16]。它通过提出严格线性相关或相关性较强的自变量的信息,选择其中某些维度来表征原有数据,以此达到降维的目的。通常,它对数据各维度进行信息贡献率的计算,并对数据维度按照贡献率排序。之后,可以根据需要自行选取特定的维度来表征原始数据。

假设问题中有 p 个指标,把这些指标看成 p 个随机变量 $X_1, X_2, \cdots, X_p$,主成分分析是要把这 p 个指标问题转化为 p 个指标的线性组合问题。这些新指标 $F_1, F_2, \cdots, F_k (k \leqslant p)$,遵循保留主要信息量原则来反映原来指标信息,并且它们相互之间独立。

$$F_1 = u_{11}X_1 + u_{21}X_2 + \ldots u_{p1}X_p$$
$$F_2 = u_{12}X_1 + u_{22}X_2 + \cdots u_{p2}X_p$$
$$\cdots\cdots$$
$$F_p = u_{1p}X_1 + u_{2p}X_2 + \cdots u_{pp}X_p$$

满足如下条件:

(1)每个主成分系数平方和是 1,即

$$u_{1i}^2 + u_{2i}^2 + \cdots + u_{pi}^2 = 1$$

(2)主成分之间相互独立,即

$$\mathrm{Cov}(F_i, F_j) = 0, i \neq j, i, j = 1, 2, \cdots, p$$

(3)主成分的方差递减,重要性递减,即

$$\mathrm{Var}(F_1) \geqslant Var(F_2) \geqslant \cdots \geqslant Var(F_p)$$

$F1$、$F2\cdots Fp$ 分别称为原始变量的第一、第二、第 p 个主成分。如图 4 所示为主成份分类散点图:

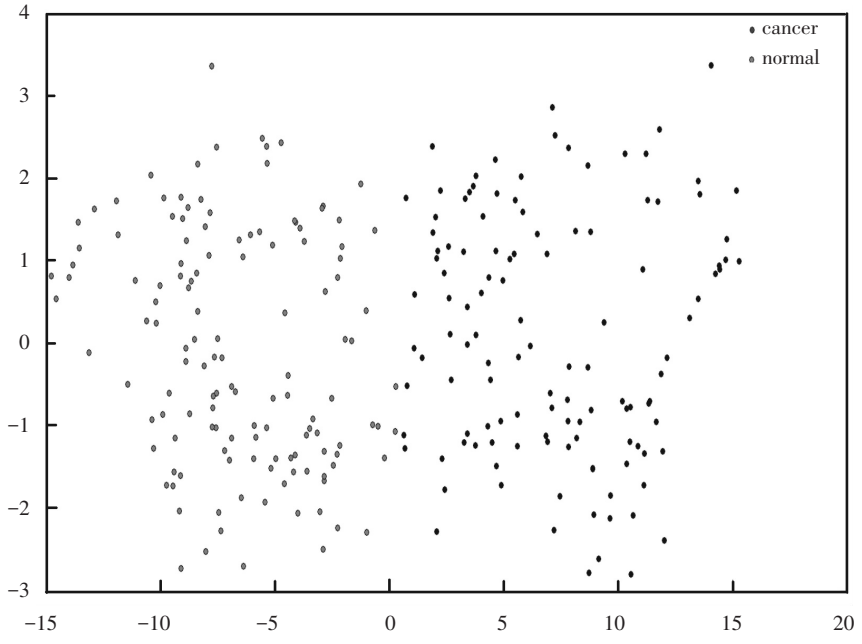

图 4 8-7-02 数据集分类结果散点图

Fig.4 Scatter of classifying 8-7-02 data set

注:黑色点为癌症数据,灰色点为正常数据。

Notes:Black dots are cancer datas, grey dots are normal datas.

1.3 支持向量机

支持向量机(Support Vector Machine,SVM)从线性可分情况下的最优分类面发展而来。最优分类面就是要求分类线不但能将两类正确分开(训练错误率为 0),且使分类间隔最大。支持向量机考虑寻找一个满足分类要求的超平面,并且使训练集中的点距离分类面尽可能的远,也就是寻找一个分类面使它两侧的空白区域(Margin)最大[17]。

1.4 K 折交叉验证

交叉验证(K-fold cross-validation)是机器学习数据重抽样常用的方法,并且被广泛使用。交叉验证主要有三种,Handout 验证、k 折交叉验证(K-fold cross-validation)、留一验证(Leave-m-out)。本文主要使用 k 折交叉验证(K-fold cross-validation)。其基本过程为:将样本集随机分为 K 个集合,通常分为 K 等份,对其中的 $K-1$ 个集合进行训练,剩下的一个集合用来在分类器中进行样本测试。该过程重复 K 次,取 K 次过程中的测试错误的平均值作为推广误差。

2　实　验

2.1　实验数据

本实验中，总共使用了三组 SELDI-TOF 蛋白质质谱数据集来测试分类器的性能。三组数据集中有一组高分辨率卵巢癌数据集、两组低分辨率卵巢癌数据集。三组数据集来源于文献[18]。这些数据在文献[18]中分别给予了命名，本论文沿用文献[18]中的命名。下面简单介绍这三组数据。

2.1.1　8-7-02 数据集

这组低分辨率卵巢癌数据集在采集数据过程中使用了 WCX2 蛋白质芯片，然后使用升级的 PBSII 型 SELDI-TOF 质谱仪来生成质谱数据。这组数据集包含 162 个卵巢癌样本和 91 个正常样本。每个样本有 15 154 个特征。

2.1.2　这组数据也是低分辨率卵巢癌数据，亦是采用 WCX2 蛋白质芯片制备样本的。这组数据集由 100 个卵巢癌样本和 100 个正常样本组成。每个样本有 15 154 个特征。

2.1.3　OvarianCD_PostQAQC 数据集

此组为高分辨率卵巢癌质谱数据集。它由 ABI Qstar 型 SELDI-TOF 质谱仪生成的非随机卵巢癌样本和正常样本组成。卵巢癌样本 121 个，正常样本 95 个。每个样本由 15 154 个特征组成。

2.2　基本思路方法

将数据预处理后，通过小波分析技术进行降维处理，之后使用 PCA 技术，继续降维，取出主成分属性。然后用支持向量机(SVM)作为分类器，通过 k-fold 交叉验证，分类数据，并评估其性能。主要过程如图 5 所示：

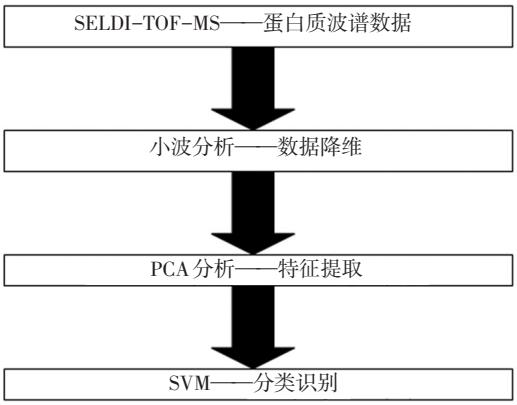

图 5　实验流程图

Fig.5　Experiment flow

2.3　实　验

对 8-7-02 数据集实验结果进行分析。实验过程中，首先，确定 PCA 分析所取最佳属性，实验中，取能表征数据集 90%以上主成分分量的最佳属性。经测试，8-7-02 数据集经过小波分析和主成分分析后，前 12 维属性贡献率之和达到 90.61%，故取其前 12 维属性，如图 6 所示：

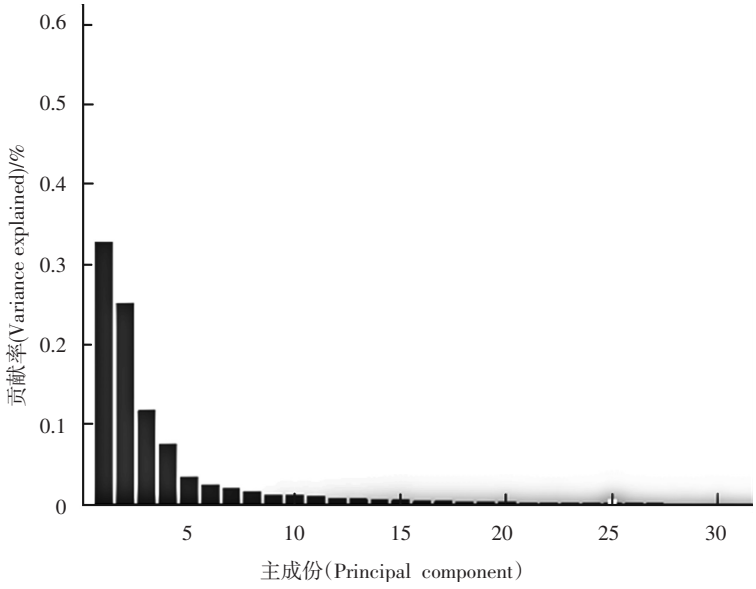

图 6　8-7-02 数据集部分主要维度属性贡献率

Fig.6　Contribution rates of some main properties of 8-7-02 data set

注：前十二维属性贡献率之和为 90.61%。

Notes：Sum of the contribution rates of the first twelve properties is 90.61%.

通过图 3 的思路，对 8-7-02 数据集进行分类，其中 k-fold 验证中参数取 5，小波变换过程中，分解层数分别取 1 到 5 层，小波基分别取 haar 和 dbN 小波系。最终得到在不同小波分解层数和不同小波基

条件下的分类情况。结果如表 1 所示：

实验一：取前十二维属性，对比分析不同小波基、分解层数实验效果。

由实验结果数据可以看到，随着小波分解层数增加，分类正确率、灵敏性、特异性都略有下降，每增加一层分解，数据属性维度就会减少一半，数据维度太多或太少，都不能很好的实现分类效果。另外，小波分解之后，对得到数据进行主成分分析，数据的前少数属性维即可很好的表征数据特征，不需要太多冗余属性维，这大大降低了数据维度，为之后的分类减轻了很大的负担，极大的提高了效率。最终经实验分析得出，8-7-02 数据集在使用 db3 小波基，小波分解层数为 1，取前 12 维属性时，其分类效果最佳：正确率 98.38%，灵敏性 98.79%，特异性98.15%。见表 1、表 2、表 3。

实验二：固定小波基和分解层数，对比选取不同主成分属性实验效果。

当分别取前 6、9、12 维属性，使用 db4 小波基、3 层分解时，实验结果对比如表 4 所示：

表 1　不同小波基在不同分解层数条件下分类正确率

Table 1　Accuracy of classification under the different conditions of wavelet basis and levels

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.983 6	0.983 8	0.983 4	0.981 8	0.979 1	0.982 6	0.983 0
2	0.981 8	0.983 5	0.980 4	0.983 6	0.978 9	0.979 6	0.982 0
3	0.977 1	0.978 3	0.973 9	0.970 9	0.969 6	0.968 6	0.975 5
4	0.969 0	0.976 5	0.970 2	0.977 3	0.974 7	0.962 1	0.969 4
5	0.956 7	0.962 6	0.962 5	0.965 8	0.957 1	0.964 0	0.957 3

表 2　不同小波基在不同分解层数条件下对应灵敏性

Table 2　Sensitivity under the different conditions of wavelet basis and levels

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.986 8	0.987 9	0.984 1	0.985 2	0.976 4	0.982 4	0.987 9
2	0.979 7	0.988 5	0.978 6	0.988 5	0.981 9	0.975 8	0.979 1
3	0.972 5	0.976 9	0.968 7	0.965 4	0.963 2	0.957 7	0.975 3
4	0.966 5	0.969 8	0.969 2	0.977 5	0.968 1	0.968 1	0.967 0
5	0.970 3	0.975 3	0.973 6	0.979 1	0.958 8	0.974 7	0.968 7

表 3　不同小波基在不同分解层数条件下对应特异性

Table 3　Specificity under the different conditions of wavelet basis and levelsc

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.981 5	0.981 5	0.981 2	0.981 2	0.979 3	0.979 6	0.981 2
2	0.982 1	0.982 1	0.981 2	0.982 4	0.978 4	0.981 2	0.984 0
3	0.975 3	0.977 2	0.980 6	0.975 6	0.970 7	0.970 7	0.976 2
4	0.967 3	0.979 6	0.969 8	0.980 9	0.975 9	0.957 7	0.968 8
5	0.947 2	0.953 7	0.960 2	0.954 0	0.959 3	0.956 8	0.952 8

表 4　不同维度数据在相同小波分解条件下结果

Table 4　Classific result under the conditions of different dimensions and samewavelet basis

特性	选择维度		
	12	9	6
贡献率	0.906 1	0.884 1	0.835 3
正确率	0.973 6	0.973 3	0.963 6
灵敏性	0.968 7	0.984 1	0.948 4
特异性	0.980 6	0.967 3	0.972 2

从表中数据，我们可以看出，随着维度数量的增加，正确率逐渐提高，但是当维度达到一定数量之后，正确率的增加量逐渐减小。

其他数据集同样经上述思路进行实验处理后，具体实验数据如下：

3.3.1　4/3/02 数据集：

经实验处理后，本组数据前 10 维属性贡献率之和达到 90.25%，分类实验取前 10 维如图 7：

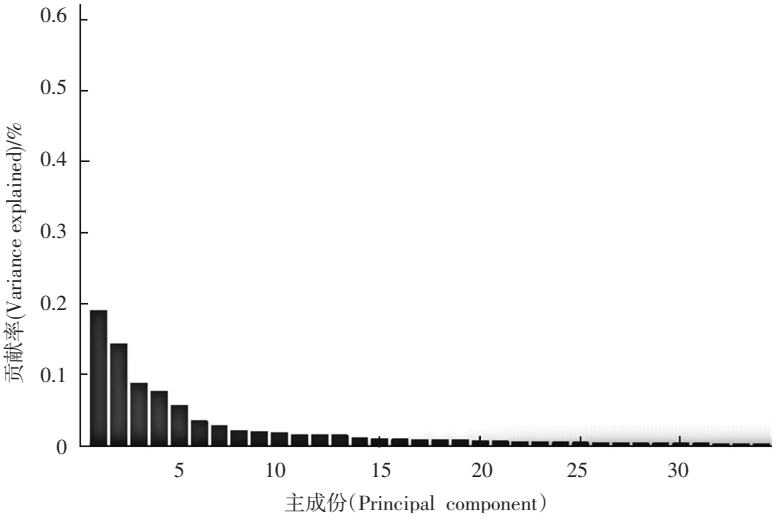

图 7　4/3/02 数据集部分主要维度属性贡献率

Fig.7　Contribution rates of some main properties of 4/3/02 data set

注:前十维属性贡献率之和为 90.25%。

Notes:Sum of the contribution rates of the first tenth properties is 90.25%.

由实验数据我们看到,对 4/3/02 数据集进行实验,当使用 db8 小波基,小波分解层数为 1 时,其分类效果最佳:正确率 86.45%,灵敏性 87.00%,特异性 85.90%。见表 5、表 6、表 7。

表 5　不同小波基在不同分解层数条件下分类正确率(4/3/02 数据集)

Table 5　Accuracy of classification under the different conditions of wavelet basis and levels(4/3/02 data set)

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.844 5	0.826 5	0.818 8	0.864 3	0.864 5	0.851 8	0.845 0
2	0.812 3	0.832 2	0.846 0	0.839 5	0.818 3	0.842 5	0.816 8
3	0.802 8	0.820 3	0.832 5	0.826 8	0.824 3	0.832 0	0.799 8
4	0.784 0	0.795 0	0.767 0	0.787 5	0.775 3	0.784 8	0.783 5
5	0.739 0	0.739 5	0.779 3	0.742 3	0.751 3	0.765 5	0.744 3

表 6　不同小波基在不同分解层数条件下对应灵敏性(4/3/02 数据集)

Table 6　Sensitivity under the different conditions of wavelet basis and levels(4/3/02 data set)

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.849 0	0.800 5	0.792 5	0.893 0	0.870 0	0.855 0	0.848 5
2	0.802 0	0.839 0	0.855 0	0.843 0	0.800 0	0.836 5	0.806 0
3	0.790 5	0.821 0	0.819 0	0.814 0	0.825 0	0.843 0	0.790 5
4	0.778 5	0.793 5	0.740 5	0.768 0	0.743 5	0.752 0	0.779 5
5	0.718 0	0.705 0	0.764 0	0.684 0	0.702 0	0.721 0	0.727 0

表 7　不同小波基在不同分解层数条件下对应特异性(4/3/02 数据集)

Table 7　Specificity under the different conditions of wavelet basis and levelsc(4/3/02 data set)

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.840 0	0.852 5	0.845 0	0.835 5	0.859 0	0.848 5	0.841 5
2	0.822 5	0.825 5	0.837 0	0.836 0	0.836 5	0.848 5	0.827 5
3	0.815 0	0.819 5	0.846 0	0.839 5	0.823 5	0.821 0	0.809 0
4	0.789 5	0.796 5	0.793 5	0.807 0	0.807 0	0.817 5	0.787 5
5	0.760 0	0.774 0	0.794 5	0.800 5	0.800 5	0.810 0	0.761 5

3.3.2 OvarianCD_PostQAQC 数据集

经实验处理后,本组数据前 145 维属性贡献率之和达到 90.14%,分类实验取前 145 维,如图 8:

由实验数据我们看到,对 OvarianCD_PostQAQC 数据集进行实验,当使用 db10 小波基,小波分解层数为 3 时,其分类效果最佳:正确率 92.18%,灵敏性 91.00%,特异性 93.10%。见表 8、表 9、表 10。

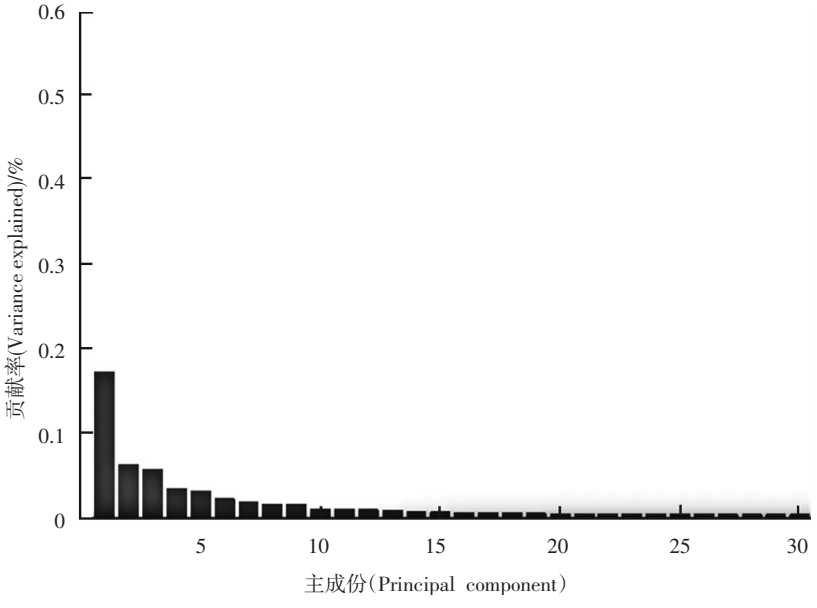

图 8 OvarianCD_PostQAQC 数据集部分主要维度属性贡献率

Fig.8 Contribution rates of some main properties of OvarianCD_PostQAQC data set:sum of the contribution rates of the first 145th properties is 90.14%

注:前 145 维属性贡献率之和为 90.14%。

Notes:Sum of the contribution rates of the first 145th properties is 90.14%.

表 8 不同小波基在不同分解层数条件下分类正确率(OvarianCD_PostQAQC 数据集)

Table 8 Accuracy of classification under the different conditions of wavelet basis and levels(OvarianCD_PostQAQC data set)

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.838 2	0.846 1	0.855 1	0.861 8	0.894 4	0.867 8	0.848 4
2	0.894 2	0.902 5	0.890 7	0.879 6	0.841 0	0.910 0	0.889 8
3	0.899 5	0.878 2	0.908 1	0.866 7	0.912 5	0.921 8	0.903 9
4	0.908 3	0.879 2	0.905 3	0.888 7	0.867 6	0.906 3	0.906 3
5	0.901 4	0.870 1	0.897 0	0.901 4	0.888 4	0.864 4	0.899 8

表 9 不同小波基在不同分解层数条件下对应灵敏性(OvarianCD_PostQAQC 数据集)

Table 9 Sensitivity under the different conditions of wavelet basis and levels(OvarianCD_PostQAQC data set)

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.795 3	0.807 9	0.819 5	0.805 8	0.840 0	0.830 5	0.810 5
2	0.869 5	0.875 3	0.862 6	0.851 6	0.805 3	0.903 7	0.863 7
3	0.893 7	0.859 5	0.894 7	0.857 4	0.903 7	0.910 0	0.899 5
4	0.874 7	0.840 0	0.888 4	0.857 4	0.840 0	0.881 1	0.873 7
5	0.897 4	0.854 2	0.887 4	0.891 6	0.892 6	0.850 5	0.890 0

表 10　不同小波基在不同分解层数条件下对应特异性(OvarianCD_PostQAQC 数据集)

Table 10　Specificity under the different conditions of wavelet basis and levelsc(OvarianCD_PostQAQC data set)

分解层数	小波基						
	db1	db3	db4	db6	db8	db10	haar
1	0.871 9	0.876 0	0.883 1	0.905 8	0.937 2	0.897 1	0.878 1
2	0.913 6	0.924 0	0.912 8	0.901 7	0.869 0	0.914 9	0.910 3
3	0.904 1	0.893 0	0.918 6	0.874 0	0.919 4	0.931 0	0.907 4
4	0.934 7	0.909 9	0.918 6	0.913 2	0.889 3	0.926 0	0.931 8
5	0.904 5	0.882 6	0.904 5	0.909 1	0.885 1	0.875 2	0.907 4

4　讨论与结论

经过一系列的实验,我们发现,同一组数据,在进行小波分解时,采用同一小波基,当分解层数不同时,分类结果会略有不同,如 8-7-02 数据集在使用 db3 小波基时,在一到五层分解时正确率分别为 98.38%、98.35%、97.83%、97.65%、96.26%。另外,不同小波基,在相同分解层数条件下,对于数据分类结果,也会有不同影响,正确率会有所不同,但是差别不大,如 8-7-02 数据集在进行 2 层小波分解时,分别使用 db1、db3、db4、db6、db8、db10、haar 小波基时正确率分别为 98.18%、98.35%、98.04%、98.36%、97.89%、97.96%、98.20%。与文献[4]的综合识别率基本持平,但是在数据处理中,通过小波分析和主成分分析大大降低了数据维度,简化了运算。

本文提出的模型中,先对蛋白质波谱数据进行小波分解,然后通过主成分分析提取特征,之后将特征送入支持向量机分类。经实验,本模型可以有效的降低数据计算量,提高效率,并能较好的对蛋白质波谱数据进行分类。

参考文献(References)

[1] 吕红. 蛋白质质谱分析法的研究进展[J]. China Pharmacy, 2010, 21(25): 2388-2389.
LÜ Hong. The study progress of protein mass spectrometry analysis[J]. China Pharmacy, 2010, 21(25): 2388-2389.

[2] 严勇, 王鑫, 杨慧中. 基于决策树与质谱分析数据的癌症判别[J]. 无锡职业技术学院学报, 2013, 12(1): 31-33.
YAN Yong, WANG Xin, YANG Huizhong. Cancer discriminant based on the decision tree and mass spectrometry analysis data[J]. Proceedings of the Wuxi Institute of Technology, 2013, 12(1): 31-33.

[3] SCHAPIRE R, FREUND Y, BARTLETT P, WEE SUN L. Boosting the margin: a new explanation for the effectiveness of voting methods [J]. The Annals of Statistics, 1988, 26(5): 1651-1686.

[4] 邹修明, 罗楠, 孙怀江. 基于T检验与支持向量机的蛋白质质谱数据分析[J]. 淮阴师范学院学报(自然科学), 2011, 10(5): 409-413.
ZOU Xiuming, LUO Nan, SUN Huaijiang. Protein mass spectrometry analysis based on T-test and svm [J]. Proceedings of the Huaiyin Normal Univerity (natural sciences), 2011, 10(5): 409-413.

[5] 杨合龙, 祝磊, 韩斌. 运用近邻传播聚类分析进行SELDI-TOF 蛋白质谱特征选择[J]. 中国生物医学工程学报, 2013, 32(1): 14-18.
YANG Helong, ZHU Lei, HAN Bin. SELDI-TOF protein mass spectrometry feature selection based on neighbor clustering analysis [J]. Chinese Journal of Biomedical Engineering, 2013, 32(1): 14-18.

[6] KUEHL B, MARTEN S, BISCHOFF Y, et al. MALDI-ToF mass spectrometry-multivariate data analysis as a tool for classification of reactivation and non-culturable states of bacteria[J]. Anal Bioanal Chem, 2011, 401: 1593-1600.

[7] EBERLIN L, NORTON I, DILL A, et al. Classifying human brain tumors by lipid imaging with mass spectrometry[J]. Cancer Research, 2012, 72: 645-654.

[8] 王昭鑫, 刘毅慧. 主元余像集主成分分析在蛋白质质谱数据中的应用[B]. 生物信息学, 2009, 7(3): 219-222.
WANG Zhaoxin, LIU Yihui. Application of 2nd PCA on protein mass spectrometry data[B]. Chinese Journal of Bioinformatics, 2009, 7(3): 219-222.

[9] BEHDAD M, FRENCH T, BARONE L, et al. On principal component analysis for high-dimensional XCSR [J]. Evolutionary Intelligence, 2012, 5(2): 129-138.

[10] Baldi P, Long A. A Bayesian framework for the analysis of microarray expression data:regularized t-test and statistical inferences of gene changes[J]. Bionformatics, 2001, 17: 509-519.

[11] ZHAO J. Asymptotic convergence of dimension reduction based boosting in classification[J]. Journal of Statistical

Planning and Inference, 2013, 143(4): 651-662.

[12] 李义峰, 刘毅慧. 基于遗传算法的蛋白质质谱数据特征选择[J]. 计算机工程, 2009, 35(19): 192-197.
LI Yifeng, LIU Yihui. Feature selection for protein mass spectrometry data based on genetic algorithm [J]. Computer Engineering, 2009, 35(19): 192-197.

[13] 李义峰, 刘毅慧. 基于模拟退火算法的高分辨率蛋白质质谱数据特征选择[J]. 生物信息学, 2009, 2(7): 85-90.
LI Yifeng, LIU Yihui. Feature selection based on simulated annealing algorithm for high-resolutio protein mass spectrometry data [J]. Chinese Journal of Bioinformatics, 2009, 2(7): 85-90.

[14] LIU Yihui. Feature extraction and dimensionality reduction for mass spectrometry data[A]. Computers in Biology and Medicine, 2009, 39: 818-823.

[15] 张德丰. MATLAB 小波分析(第二版)[M]. 北京: 机械工业出版社, 2011.
ZHANG Defeng. The wavelet analysis of matlab (the second edition) [M]. Beijing: China Machine Press, 2011.

[16] GELADI P. Notes on the history and nature of partial least squares (PLS) modeling[J]. Journal of Chemometrics, 1988, 2: 231-246.

[17] 边肇祺, 张学工. 模式识别(第二版)[M]. 北京:清华大学出版社, 2003.
BIAN Zhaoqi, ZHANG Xuegong. Pattern recognition (the second edition)[M]. Beijing: Tsinghua University Press, 2003.

[18] 李义峰. 基于优化算法的蛋白质质谱数据分析[D]. 济南:山东轻工业学院, 2009.
LI Yifeng. Optimization algorithms based protein mass spectrometry data analysis [D]. Jinan: Shandong Polytechnic University, 2009.