基于 SVM 的非相关线性判别分析算法研究

张小丹,吕建平
ZHANG Xiao-dan,LV Jian-ping

苏州大学 电子信息学院,江苏 苏州 215021
School of Electronics and Information Engineering of Soochow University,Suzhou,Jiangsu 215021,China
E-mail:xiaodan.albert@gmail.com

ZHANG Xiao-dan,LV Jian-ping.Research of uncorrelated linear discriminant analysis based on support vector machine. Computer Engineering and Applications,2008,44(4):227-229.

Abstract: The classification of tissue samples based on gene expression data is an important problem in medical diagnosis of diseases.In gene expression data,the number of genes is usually very high (in the thousands) compared to the number of data samples(in the tens);that is,the data dimension is large compared to the number of data points(such is undersampled problem). Too high dimension (the number of features or genes) makes the task of classification quite challenging.This paper presents that ULDA and SVM are combined to classify colon tissue samples.Compared to other methods,the effect of classification is improved,the results prove the feasibility and effectiveness of this method.

Key words: Uncorrelated Linear Discriminant Analysis;SVM;gene expression profiling;classification

摘 要:基于基因表达谱对组织样本进行分类,在疾病诊断领域,是个非常重要的研究课题。在基因表达数据中,基因的数量(几千个)相对于数据样本(几十个)的个数通常比较多;也就是说,数据的维数相比于数据点的个数来说比较高(这个就是采样不足问题)。过高的维数(特征或基因数)将给分类问题带来极大的挑战。提出了结合非相关线性判别式分析方法(ULDA)和支持向量机(SVM)分类算法,对结肠癌组织样本进行分类识别,并同其他方法作了比较研究,分类效果得到了提高;结果表明了该方法的可行性和有效性。

关键词:非相关线性判别分析;支持向量机;基因表达谱;分类
文章编号:1002-8331(2008)04-0227-03 **文献标识码**:A **中图分类号**:TP391.4

1 引言

DNA 微阵列是一项新的技术,用来在单次实验中检测成千上万个基因的表达水平。对一些小的有机体的整个基因组中的大量基因进行表达水平的检测后,提出根据基因表达水平描述细胞单元特性,也就是基于基因表达水平确定细胞单元的状态和功能。最基本的描述问题是识别一组基因和它们的表达模式来刻画某一个细胞单元的状态或预测未来的状态。这个研究方向中最关键的一步是,根据它们的基因表达水平,利用工具对组织样本进行分类。特别是给定一个基因表达数据集,事先依据疾病类型分成了几类,目标是判决一个新的组织样本可能属于哪一类。文献[3,5]中有进一步的讨论。

基因表达数据具有一些特性使得分类研究相当具有挑战性。表达数据通常包含大量的基因(典型地,几千个)和少量的样本(通常,几十或一两百个)。在机器学习领域中,这些数据被认为是具有高的维数和小的样本数,属于采样不足数据集。目前,已经提出许多方法来降低数据维数,也就是通过选择一个包含大多数有关基因的子集,利用选出的基因(特征)子集来对组织样本进行分类,既有利于提高分类精度,又能加快分类速度。本文依据前人提出的非相关线性判别式分析算法(ULDA)和支持向量机理论(SVM),把这两种方法结合起来,加以改进,用于前列腺癌的基因表达谱数据进行分类实验,取得了良好的分类结果,并和已有的分类方法比较,实现了算法的速度和精度的提高。

2 实验数据描述和预处理

结肠癌数据集来自于文献[2]的分析,它包括 40 个肿瘤组织样本和 22 个正常组织样本,每个样本均有 2000 基因表达谱数据。该数据集可从 http://microarray.princeton.edu/oncology 网站获得。实验中,首先对原始数据进行归一化预处理,使得每个基因表达值的均值为 0,方差为 1,然后随机地将数据集分成训练集和独立测试集,训练集占整个数据集的三分之二,而测试集占三分之一,并且这两个集中正常和肿瘤样本数目的比例都是近似。为了降低不确定性,重复 50 次进行训练集和测试集的划分,得到的精度是个平均值。如图 1 所示。

3 非相关线性判别式分析的特征提取和维数降低

线性判别式分析(LDA)[12]是一个用于特征提取和维数降

作者简介:吕建平,男,副教授,苏州大学电子信息学院,研究方向:生物医学信息,模式识别等;张小丹(1981-),男,苏州大学电子信息学院研究生,研究方向:模式识别。

收稿日期:2007-06-01 **修回日期**:2007-08-09

训练集
正常样本:14
肿瘤样本:26

测试集
正常样本:8
肿瘤样本:14

图 1　结肠癌数据样本集

低的很有效的方法。它已经被广泛地应用于许多领域,例如,人脸识别[4],文本分类,微阵列数据分类[8]等。假定一个高维的数据集事先已分成几类,经典的 LDA 在于找到一个最优转换,把数据映射到低维空间(并且保留分类结构),目标是最小化类内距离,同时最大化类间距离,达到最大判别。最优转换可以通过对离散矩阵进行特征分解计算出来。而经典 LDA 的一个内在局限性是目标函数要求至少其中的一个离散矩阵是非奇异的。然而,在许多应用中,例如微阵列数据分类、人脸识别、文本分类,所有的离散矩阵可能是奇异的, 因为数据来自非常高维的空间,而且通常维数超过样本数目;这就是采样不足或奇异问题。非相关线性判别式分析(ULDA)是一种基于判别式分析的特征提取和维数降低算法,这个方法有几个优点:第一,通过 ULDA 获得的转换之后的特征是原始特征(基因)的线性组合。换句话说,ULDA 考虑了不同基因之间的关系;第二,理论证明了在转换后的空间中,特征是非相关的,这样就保证了在降维空间中特征之间的最小冗余;第三,大家知道,经典判别式分析由于类内离散矩阵是奇异的[12]而不适用。对于基因表达数据,基因数远大于样本数,类内离散矩阵通常是奇异的。ULDA 应用广义奇异值分解来克服采样不足问题。

给定一个基因表达数据矩阵 $\boldsymbol{A}=(a_{ij})\in\mathbb{R}^{p\times n}$,其中每一列对应一个样本,每一行对应一个特定的基因,试图找到一个线性转换 $\boldsymbol{G}\in\mathbb{R}^{p\times l}(l<p)$,把 $\boldsymbol{A}$ 的每一列 a_i,其中 $1\leqslant i\leqslant n$,从 p 维空间映射到 l 维空间中的 z_i。即

$$\boldsymbol{G}:a_i\in\mathbb{R}^p\rightarrow z_i=\boldsymbol{G}^{\mathrm{T}}\cdot a_i\in\mathbb{R}^l \tag{1}$$

转换后得到的数据矩阵为 $\boldsymbol{A}^L=\boldsymbol{G}^T\cdot\boldsymbol{A}\in\mathbb{R}^{l\times n}$,包含 l 行,也就是在降维空间中每一个样本有 l 个特征。降维空间中的每一个特征是原始高维空间中特征的线性组合,线性组合的系数依赖于转换矩阵 $\boldsymbol{G}$。

假设数据集中有 k 个类别,m_i,S_i,P_i 分别表示第 i 个类的质心,协方差矩阵和先验概率,m 是整个数据集的质心。类间离散矩阵 $\boldsymbol{S}_b$,类内离散矩阵 $\boldsymbol{S}_w$,总离散矩阵 $\boldsymbol{S}_t$ 如下定义:

$$\boldsymbol{S}_b=\sum_{i=1}^{k}P_i(m_i-m)(m_i-m)^{\mathrm{T}} \tag{2}$$

$$\boldsymbol{S}_w=\frac{1}{n}\sum_{i=1}^{k}S_i \tag{3}$$

$$\boldsymbol{S}_t=\boldsymbol{S}_b+\boldsymbol{S}_w \tag{4}$$

第 i 类的协方差矩阵 $\boldsymbol{S}_i$ 可以分解为 $\boldsymbol{S}_i=\widetilde{\boldsymbol{A}}_i\widetilde{\boldsymbol{A}}_i^{\mathrm{T}}$, $\widetilde{\boldsymbol{A}}_i$ 的每一列对应于第 i 类中的一个已经减去质心的数据点。

定义如下矩阵:

$$\boldsymbol{H}_w=\frac{1}{\sqrt{n}}[\widetilde{\boldsymbol{A}}_i,\cdots,\widetilde{\boldsymbol{A}}_k]$$

$$\boldsymbol{H}_b=[\sqrt{P_1}(m_1-m),\cdots,\sqrt{P_k}(m_k-m)] \tag{5}$$

在线性代数中,同时对两个矩阵进行对角化的一个常见的方法是运用广义奇异值分解(GSVD)。假设 $\Gamma=\begin{pmatrix}\boldsymbol{H}_b^{\mathrm{T}}\\ \boldsymbol{H}_w^{\mathrm{T}}\end{pmatrix}$,这是一个 $(n+k)\times p$ 阶矩阵。通过广义奇异值分解,得到正交矩阵 $\boldsymbol{U}\in\mathbb{R}^{k\times k}$,$\boldsymbol{V}\in\mathbb{R}^{n\times n}$ 和一个非奇异矩阵 $\boldsymbol{X}\in\mathbb{R}^{p\times p}$,使得

$$\begin{pmatrix}\boldsymbol{U}&0\\0&\boldsymbol{V}\end{pmatrix}^{\mathrm{T}}\Gamma\boldsymbol{X}=\begin{pmatrix}\sum_1&0\\ \sum_2&0\end{pmatrix} \tag{6}$$

选取上述公式(6)中矩阵 $\boldsymbol{X}$ 的前 $q=\mathrm{rank}(\boldsymbol{H}_b)$ 个列作为转换矩阵 $\boldsymbol{G}$,并且经过转换后的样本的特征之间是非相关的。

4　支持向量机对维数降低后的数据进行分类

支持向量机(SVM)是由 Vapnik 等人基于统计学习理论,采用结构风险最小化原理提出的一种新的机器学习算法。通过调整判别函数使得它最好地利用边界样本点的分类信息,构造出最佳分类超平面。其主要优点有:(1)是专门针对有限样本情况的,其目标是得到现有信息下的最优解而不仅仅是样本数趋于无穷大时的最优值;(2)算法最终将转化成为一个二次型寻优问题,从理论上说,得到的将是全局最优点,解决了在神经网络方法中无法避免的局部极值问题;(3)算法将通过非线性变换将原数据空间转换到高维的特征空间,在高维空间中构造线性判别函数来实现原空间中的非线性判别函数,这种特殊性质能保证机器有较好的推广能力,同时它巧妙地解决了维数问题。

若给定的样本集为 (x_i,y_i),$i=1,\cdots,n$,$x_i\in R^d$,$y_i\in\{-1,+1\}$ 是类别标号,则支持向量机的判别函数为:

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{sv}\alpha_i^*y_iK(x_i,x)+b^*\right) \tag{7}$$

式中 sv 是支持向量的个数,$K(x_i,x)$ 为核函数,本文支持向量机核函数选用 RBF 函数:

$$K(x_i,x)=e^{-\frac{\|x-x_i\|^2}{\sigma^2}} \tag{8}$$

在样本识别中,为获得对分类错误率的可靠估计,本文采用如下两个步骤进行样本识别:

(1)在训练集上采用“留一交叉检验”(Leave-One-Out Cross Validation,LOOCV)的方　法进行样本类型的识别;

(2)以训练集中的所有样本作为 SVM 的训练样本,对独立测试集中的样本逐一进行识别,该过程称为“独立测试”实验。

实验的整个流程如图 2。

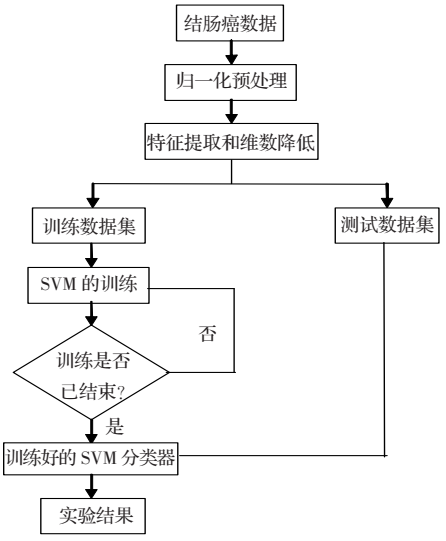

图 2　分类流程图

5　实验过程和结果分析

本文先对结肠癌数据进行归一化预处理,即计算每一个基

因的均值和标准偏差，并使基因的每个值减去这个均值后再除以标准偏差；接着用非相关线性判别式分析算法（运用广义奇异值分解）对基因表达数据进行特征提取和维数降低；最后用 RBF 支持向量机对数据进行分类训练，构造 SVM 分类器，并对测试数据进行分类识别，以分类精度作为评判分类器性能的标准。实验中，本文取 σ=10，C=400 的 RBF 核函数，用 Matlab 进行程序的编写。并与其他几种方法进行了比较。结果如表 1。

从表 1 中可以看出，由于结肠癌数据集中的 2 000 个基因已经是经过了很好处理后的结果，所以 KNN（K-近邻）、DLDA（对角线性判别式分析）、Fisher 分类器，以及直接用 SVM 进行分类，得到的准确率都不是很高。但可以看出本文提出的方法得到的精度最高，说明该方法的有效性和可行性。

表 1　五种分类器的分类精度的比较结果

分类方法	ULDA+SVM	KNN	DLDA	SVM	Fisher
分类精度	87.23%	82.38%	86.29&	85.05%	87.12%

6　小结

针对肿瘤基因表达数据维数高、样本少，即采样不足问题。本文提出的结合方法有效地解决了维数问题，并剔除了冗余信息，使得特征之间具有非相关性；并使用 SVM 分类器进行了分类识别，提高了分类精度，实现了方法的改进，表明了该方法的有效性。

目前微阵列技术产生了大量的基因表达数据，能否将提出的方法用于这些数据集，并同样得到好的结果，即提高泛化能力，而且要提高分类效率，是进一步研究的方向。

参考文献：

[1] Ward J J，McGuffin L J，Buxton B F，et al.Secondary structure prediction with support vector machines[J].Bioinformatics，2003，19(13)：1650-1655.

[2] Alon U，Barkai N，Notterman D A，et al.Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide Arrays[C]//Proc Nat'l Academy of Science，1999，96：6745-6750.

[3] Ben-Dor A，Bruhn L，Friedman N，et al.Tissue classification with gene expression profiles[J].J Computational Biology，2000，7：559-584.

[4] Belhumeour P N，Hespanha J P，Kriegman D J.Eigenfaces vs.Fisherfaces：recognition using class specific linear projection [J].IEEE Trans Pattern Analysis and Machine Intelligence，1997，19(7)：711-720.

[5] Brown M P S，Grundy W N，Lin D，et al.Knowledge-based analysis of microarray gene expression Data by Using Support Vector Machines[J].Proc Nat'l Academy of Science，2000，97：262-267.

[6] Zhao Y，Pinilla C，Valmori D，et al.Application of support vector machines for T-cell epitopes prediction[J].Bioinformatics，2003，19(15)：1978-1984.

[7] Golub G H，Van Loan C F.Matrix Computations[M].[S.l.]：Johns Hopkins Univ Press，1996.

[8] Dudoit S，Fridlyand J，Speed T P.Comparison of discrimination methods for the classification of tumors using gene expression data[J].J Am Statistical Assoc，2002，97：77-87.

[9] Howland P，Jeon M，Park H.Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition[J].SIAM J Matrix Analysis Applications，2003，25(1)：165-179.

[10] Furey T S，Cristianini N，Duffy N，et al.Support vector machine classification and validation of cancer tissue samples using microarray expression data[J].Bioinformatics，2000，16：906-914.

[11] 李泽，包雷，黄英武，等.基于基因表达谱的肿瘤分型和特征基因的选取[J].生物物理学报，2002，18(4)：413-417.

[12] Fukunaga K.Introduction to statistical pattern recognition[M].[S.l.]：Academic Press，1990.

[13] 邓乃扬，田英杰.数据挖掘中的新方法[M].北京：科学出版社，2004.

（上接 226 页）

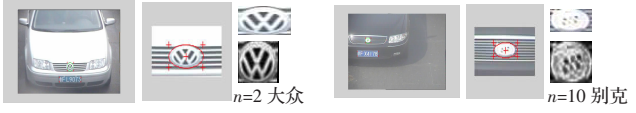

图 6　部分车标定位与识别结果

5　结论

利用车牌和车标的位置关系，可以确定车标的大致区域，利用垂直边缘算子 E 可以剔除水平格栅等噪声影响，并可精确定位并切割出车标图像；利用车标图像的矩形和椭圆两种表示形式的 4 个切点和 1 个中心点进行车标配准；为减少车标误识率，利用 PCA 进行车标图像的重构，与车标原始图像之间建立的似真度函数可以将非车标区排除；最后利用边缘不变矩对已确认的真实车标与标准车标进行欧几里德距离计算，根据其最小距离进行车标识别。实验结果表明，该方法具有很好的鲁棒性和工程实用性。后续工作可对车灯区域进行下一步识别工作。

参考文献：

[1] Wayne W，Justin W，Eric K，et al.Development of class models for model-based automatic target recognition[C]//Proceedings of SPIE：The International Society for Optical Engineering，1999，3721：650-660.

[2] Wu W.An automatic system of vehicle number-plate recognition based on neural networks[J].Journal of Systems Engineering and Electronics，2001，12(2)：63-72.

[3] 刘怡光，游志胜.一种用于图像目标识别的神经网络及其在车型识别应用[J].计算机工程，2003，29(3)：30-32.

[4] 李贵俊，刘正熙，游志胜，等.基于能量增强和形态学滤波的车标定位方法[J].光电子.激光，2005，16(1)：76-79.

[5] 庄永，杨红雨，游志胜，等.一种快速车标定位方法[J].四川大学学报：自然科学版，2004，41(6)：1167-1171.

[6] 罗彬，游志胜，曹刚.基于边缘直方图的快速车辆标志识别方法[J].计算机应用研究，2004.21(6)：150-157.

[7] Zheng D，Zhao Y，Wang J.An efficient method of license plate location[J].Pattern Recognition Letters，2005，26(15)：2431-2438.

[8] Nelson L J.License plate recognition[J].Advanced Imaging，2002，17(9)：28-30.

[9] 李文举，梁德群，张旗.基于边缘颜色对的车牌定位新方法[J].计算机学报，2004，27(2)：204-208.

[10] 王枚，王国宏.利用伴生与互补的颜色特征的车牌定位新方法[J].计算机工程与应用，2007，43(1)：206-209.

[11] Rafael C G，Richard E W，Steven L E.Digital image processing using MATLAB[M].Beijing：Publishing House of Electronics Industry，2005，7：484.

[12] Hu M K.Visual pattern recognition by moment invariants[J].IRE Transactions on Information Theory，1962，8：179-187.