文章编号:1000-582X(2011)12-102-07

降维近似支持向量机基因芯片数据分类器

王立鹏[a],袁占亭[a],周智芳[b]
(兰州理工大学 a.电气工程与信息工程学院;b.石油化工学院,甘肃 兰州 730050)

摘 要:基因芯片技术的出现改变了生物医学研究的前景,其产生的海量数据是限制其发展的瓶颈问题。论文针对基因芯片数据量大、样本数低和基因维数高的特点,提出了一种对基因芯片数据进行分类的降维近似支持向量机 DRPSVM 基因芯片数据分类器。DRPSVM 采用降维的二次规划算法,使得该算法的时间复杂度和空间复杂度比传统的 PSVM 算法均有降低。通过在 CAMDA2000、colon 1 dataset 和 colon 2 dataset 等基因芯片数据集上的与 BP、Nearest、RBF、SVM 分类器的分类性能比较,DRPSVM 在数据样本少、数据维数急剧升高时,分类性能稳定、存在唯一的最优解、训练时间快,适合基因芯片数据分类的应用环境。

关键词:生物信息学;基因芯片数据;近似支持向量机;降维;分类器;二次规划
中图分类号:Q811.4,Q789 **文献标志码**:A

Microarray data classifier with dimensionality reduction proximal support vector machines

WANG Li-peng[a], *YUAN Zhan-ting*[a], *ZHOU Zhi-fang*[b]
(a. College of Electrical and Information Engineering; b. College of Petrochemical Technology, Lanzhou University of Technology, Lanzhou 730050, Gansu, P. R. China)

Abstract: DNA microarray technologies have changed the foreground of biological medicine, while the generated plentiful data is the key problem for the application of microarrays. Microarray data have the characteristics of large quantity, low sample size and high gene dimensionality. A microarray data classifier with dimensionality reduction proximal support vector machines (DRPSVM). A dimensionality reduction quadratic programming algorithm is used in DRPSVM, which shows faster training speed and smaller memory requirements than traditional PSVM does. Using CAMDA2000, colon 1 dataset and colon 2 dataset as the experimental datasets, the classification performance of DRPSVM is compared with those of BP, Nearest, RBF and SVM. DRSVM shows stable classification performance, existing one and only optimal solution and fast training which is suitable for DNA microarray data classification applications.

Key words: bioinformatics; microarray data; proximal support vector machines; dimensionality reduction; classifier; quadratic programming

---

收稿日期:2011-07-20
基金项目:国家十一五科技支撑资助项目(2006BAF01A21);甘肃自然科学基金资助项目(3ZS062-B25-037)
作者简介:王立鹏(1979-),男,兰州理工大学博士,主要从事生物信息学方向研究,(Tel)13919270774;
(E-mail) wlp_cn @hotmail.com。

生物信息学是一门新兴的交叉学科，在近 10 年中发展非常迅速，其运用计算机科学、数理学科的知识和技术服务于生物学数据研究，同时也为传统的数理研究开辟了新的研究方向[1]。基因芯片技术始于上世纪 90 年代，代表生物技术的前沿发展，从基础上改善了研究生物技术的方法和效率，对基因组学及后基因组研究产生了重要的影响[2]。海量信息的获得也对数据的分析及信息特征提取提出了新的挑战，如基因芯片数据的标准化、样本间距离的度量、样本（或基因）的监督、非监督、半监督分类等方法。这些探索基因功能的新技术、新方法[3]和新型分析工具[4]力图将信息数据和有机的生命活动结合起来，用以解释生命特征和基因功能[5]。

基因芯片数据分析技术主要有基因芯片数据分析的非监督算法、监督算法、数据挖掘、信息融合法等[6]。基因芯片数据分析的非监督算法基于聚类技术，主要包括系统聚类[7]、分割聚类[8]，自组织特征映射[9]、模糊聚类[10]、主成分分析[11]等。基因芯片数据分析的监督算法基于分类技术，主要包括线性判别分析[12-13]、k 最近邻分类法、决策树算法、神经网络、贝叶斯分类和贝叶斯网络[14-17]、支持向量机等。

其中支持向量机（SVM）由于其对小样本情况有较好分类性能，且理论成熟，已经有较多的研究将其用于基因芯片数据分析。如：支持向量机可通过训练分类器的方法去识别与已知的共调控基因表达类型相似的新基因[18]，该法较为有效地平衡了计算和学习理论的矛盾[19]。文献[20]在 SVM 中使用了柔性边际（soft margin），用以解决训练样本含有误分类基因的问题。文献[21]主要研究肾母细胞瘤复发的基因表达谱模型的发现方法，通过 SVM 在基因表达谱数据中的分析，发现了一小类可能会用于肿瘤预诊的基因。文献[22]将基于递归特征消除（RFE）技术的 SVM 用于癌变基因选择，获得了较高的准确性。但是 SVM 用于基因分析时也存在一些固有缺陷，当数据量增大、特别是数据维数急剧升高时，SVM 算法可能无法在特征空间中找到一个超平面实现完全的分割，原因可能是该核心函数并不适于此类基因的分类，或训练样本数据未经合适的特征选取，同时，训练效率也会严重下滑。

为了当基因数据维数急剧升高时，维持较高的分类准确性和效率，论文提出了一种针对基因芯片数据进行分类的降维近似支持向量机（dimensionality reduction proximal support vctor machines，DRPSVM）基因芯片数据分类器，DRPSVM 采用了降维的二次规划算法，不但能将基因数据的分类问题归结为仅含线性等式约束的二次规划问题，同时还在传统近似支持向量机（proximal support vctor machines，PSVM）的基础上维持了较好的分类准确性，并降低了分类处理的时空复杂度。

1　对分类模型的评价

分类可用数据库术语描述为：给定一训练数据的集合 T，T 中的元素记录由若干个属性描述。在属性中有且仅有一个属性作为类别属性。属性集合用矢量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$ 表示，其中 X_i 对应各非类别属性，可以具有不同的值域。当一个属性的值域为连续域时，称该属性为连续属性（numerical attribute），否则称为离散属性（discrete attribute）。用 C 表示类别属性，$C=\{C_1,\cdots,C_k\}$，即数据集有 K 个不同的类别。那么，T 就隐含地确定了一个从矢量 $\boldsymbol{X}$ 到类别属性 C 的映射函数 $H:f(X)\rightarrow C$，分类的目的就是采用某种方法（模型）将该隐含的函数 H 表示出来。

分类模型可以直接使用训练样本的分类准确率来估计，但由于一般的学习算法都有过拟合训练样本的倾向，即对产生分类模型的训练样本有很好的分类准确率，但对新样本的分类效果却比较差。特别当训练样本数量过少或存在噪声的时候，都容易导致过拟合问题的产生。因此，仅用训练样本的分类准确率评估还不能完全反映出分类模型性能的优劣。

论文采用保持（holdout）评估法评估分类模型。给定的数据集随机划分为 2 个独立部分：一个作为训练集；另一个作为测试集。通常训练集占 2/3，测试集占 1/3。利用训练集导出分类模型，再以分类模型对测试集的分类准确率来评估分类模型，见图 1。

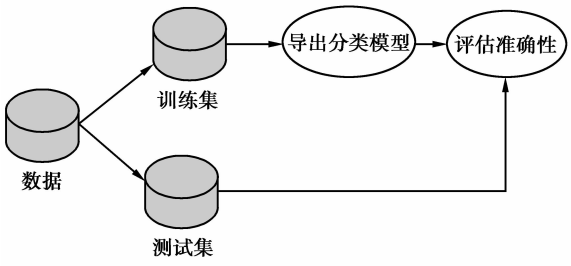

图 1　保持评估法

2 DRPSVM基因芯片数据分类器

2.1 近似支持向量机

PSVM与SVM的主要区别在于它们对应分类问题的约束条件不同，即SVM将分类归结成为线性不等式约束的二次规划问题，而PSVM将分类归结成为仅含线性等式约束的二次规划问题。

标准SVM使用n维向量空间的超平面$w\cdot x+b=0$来划分正类、负类，其分类函数为

$$c(x)=\begin{cases}+1,\text{if } w\cdot x+b\geqslant 0;\\ -1,\text{if } w\cdot x+b<0,\end{cases}\tag{1}$$

其中，分割超平面由参数w和b决定。标准SVM通过求解下面的优化问题来确定w和b

$$\begin{cases}\min & \frac{1}{2}\|w\|^2+\sum_{i=1}^{m}\xi_i;\\ \text{s.t.} & y_i(w\cdot x_i+b)+\xi_i\geqslant 1,\xi_i\geqslant 0。\end{cases}\tag{2}$$

而PSVM使用超平面$w\cdot x+b=0$划分正类、负类，但w和b通过求解另一个优化问题决定

$$\begin{cases}\min & \frac{1}{2}(\|w\|^2+b^2)+C\sum_{i=1}^{m}\xi_i^2;\\ \text{s.t.} & y_i(w\cdot x_i+b)+\xi_i=1。\end{cases}\tag{3}$$

通过对比，式(2)和(3)的主要差别在于：式(2)使用不等式约束条件，而式(3)使用等式约束条件。这表明：在SVM中，只有位于2个超平面$w\cdot x+b=1$和$w\cdot x+b=-1$之间的点才会产生训练误差，而PSVM中位于这2个超平面之内和之外的点都可能产生分类误差。因此，训练误差ξ_i可正可负，所以在式(3)的目标函数中使用了$\sum_{i=1}^{n}\xi_i^2$作为损失函数。PSVM的分类目标总结为：使正类尽量靠近超平面$w\cdot x+b=1$，负类尽量靠近超平面$w\cdot x+b=-1$，而2个超平面之间的间隔应尽量大。

2.2 降维二次规划算法

根据式(3)，PSVM的训练和学习可归为一个线性等式约束的二次规划问题。下面介绍一般等式约束问题的降维$K-T$条件

$$(\text{ECP})\begin{cases}\min & f(x);\\ \text{s.t.} & h(x)=0,\end{cases}\tag{4}$$

$f:R^n\to R,h:R^n\to R^m,m\leqslant n,P=n-m$。

$$\boldsymbol{P}(x)=\left(\frac{\partial f(x)}{\partial x_1},\frac{\partial f(x)}{\partial x_2},\cdots,\frac{\partial f(x)}{\partial x_P}\right)^{\mathrm{T}},$$

$$\boldsymbol{Q}(x)=\left(\frac{\partial f(x)}{\partial x_{P+1}},\frac{\partial f(x)}{\partial x_{P+2}},\cdots,\frac{\partial f(x)}{\partial x_n}\right)^{\mathrm{T}},$$

$$\boldsymbol{N}(x)=\begin{pmatrix}\frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_1(x)}{\partial x_2}\cdots & \frac{\partial h_1(x)}{\partial x_P}\\ \frac{\partial h_2(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_2}\cdots & \frac{\partial h_2(x)}{\partial x_P}\\ \vdots & \vdots & \vdots\\ \frac{\partial h_m(x)}{\partial x_1} & \frac{\partial h_m(x)}{\partial x_2}\cdots & \frac{\partial h_m(x)}{\partial x_P}\end{pmatrix},$$

$$\boldsymbol{M}(x)=\begin{pmatrix}\frac{\partial h_1(x)}{\partial x_{P+1}} & \frac{\partial h_1(x)}{\partial x_{P+2}}\cdots & \frac{\partial h_1(x)}{\partial x_n}\\ \frac{\partial h_2(x)}{\partial x_{P+1}} & \frac{\partial h_2(x)}{\partial x_{P+2}}\cdots & \frac{\partial h_2(x)}{\partial x_n}\\ \vdots & \vdots & \vdots\\ \frac{\partial h_m(x)}{\partial x_{P+1}} & \frac{\partial h_m(x)}{\partial x_{P+2}}\cdots & \frac{\partial h_m(x)}{\partial x_n}\end{pmatrix}。$$

定理1：设$x^0\in R^n$是ECP的最优解，f、h连续可微；若矩阵$\boldsymbol{M}(x^0)$非奇异，则

$$\boldsymbol{P}(x^0)=[\boldsymbol{M}(x^0)^{-1}\boldsymbol{N}(x^0)]^{\mathrm{T}}\boldsymbol{Q}(x^0)。\tag{5}$$

定理2：假定$h(x^0)=0$；矩阵$\boldsymbol{M}(x^0)$非奇异；$\boldsymbol{P}(x^0)=[\boldsymbol{M}(x^0)^{-1}\boldsymbol{N}(x^0)]^{\mathrm{T}}\boldsymbol{Q}(x^0)$，则$x^0$是ECP的一个$K-T$点。

推论1：如果$x^0\in\mathbf{R}^n$是方程组

$$\boldsymbol{P}(x)=[\boldsymbol{M}(x)^{-1}\boldsymbol{N}(x)]^{\mathrm{T}}\boldsymbol{Q}(x)。\tag{6}$$

的解，使得$\boldsymbol{M}(x^0)$非奇异；则x^0是ECP的一个$K-T$点。

相对经典K-T条件，由于式(6)不含Lagrange乘子，方程的维数可降低m维(等式约束个数)，因此称式(6)为ECP的降维K-T条件。

再考虑线性等式约束的二次规划问题：

$$(\text{EQP})\begin{cases}\min & f(x)=\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}G\boldsymbol{x}+\boldsymbol{g}^{\mathrm{T}}x;\\ \text{s.t.} & \boldsymbol{A}x=b,\end{cases}\tag{7}$$

其中$\boldsymbol{G}$是n阶对称矩阵，$g\in R^n$，$\boldsymbol{A}$为$m\times n$矩阵，$b\in R^m$，假定秩$\boldsymbol{A}=m$，则有

$$\nabla f(x)=\boldsymbol{G}x+\boldsymbol{g},\nabla(\boldsymbol{A}x-b)=\boldsymbol{A}。\tag{8}$$

由降维$K-T$条件知EQP的$K-T$点可由方程组：

$$\begin{cases}\boldsymbol{P}(x)=[\boldsymbol{M}^{-1}\boldsymbol{N}]^{\mathrm{T}}Q(x);\\ \boldsymbol{A}x-b=0,\end{cases}\tag{9}$$

记$p=n-m$，$\boldsymbol{M}$是$\boldsymbol{A}$中的m阶非奇异矩阵，R表示$\boldsymbol{M}$在$\boldsymbol{A}$中的列号集合，即$R=(i_1,i_2,\cdots,i_m)$；$\boldsymbol{N}$是$\boldsymbol{A}$中剩余列组成的$m\times p$矩阵，S表示$\boldsymbol{N}$在$\boldsymbol{A}$中的列号集合，即$S=(j_1,j_2,\cdots,j_p)$。相应的x可分块成$\begin{bmatrix}x_R\\ x_S\end{bmatrix}$；而

$$\boldsymbol{P}(x)=\nabla_{x_S}f(x)=\left(\frac{\partial f(x)}{\partial x_{j_1}},\frac{\partial f(x)}{\partial x_{j_2}},\cdots,\frac{\partial f(x)}{\partial x_{j_p}}\right)^{\mathrm{T}},$$

$$\boldsymbol{Q}(x)=\nabla_{x_R}f(x)=\left(\frac{\partial f(x)}{\partial x_{i_1}},\frac{\partial f(x)}{\partial x_{i_2}},\cdots,\frac{\partial f(x)}{\partial x_{i_m}}\right)^{\mathrm{T}}。$$

满足方程组(9)的 x 就是 EQP 的 K－T 点。为便于求解方程组(9),观察 EQP 的特殊性有

$$\begin{cases}\boldsymbol{P}(x)=\boldsymbol{G}_S x+\boldsymbol{g}_S,\\ \boldsymbol{Q}(x)=\boldsymbol{G}_R x+\boldsymbol{g}_R。\end{cases}\tag{10}$$

其中 $\boldsymbol{G}_s$ 是 $\boldsymbol{G}$ 取 $j_1,j_2,\cdots,j_p$ 行形成的 $p\times n$ 矩阵,$\boldsymbol{G}_R$ 是 $\boldsymbol{G}$ 取 $i_1,i_2,\cdots,i_m$ 行形成的 $m\times n$ 矩阵,$\boldsymbol{g}_s$ 是 $\boldsymbol{g}$ 取 $j_1,j_2,\cdots,j_p$ 行形成的 p 维向量,$\boldsymbol{g}_R$ 是 $\boldsymbol{g}$ 取 $i_1,i_2,\cdots,i_m$ 行形成的 m 维向量,则得到下面的方程组

$$\begin{cases}(\boldsymbol{G}_s-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_R)\boldsymbol{x}=\boldsymbol{g}_s-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{g}_R;\\ \boldsymbol{A}x=b。\end{cases}\tag{11}$$

对上述方程组分块,得到方程组

$$\begin{pmatrix}\boldsymbol{G}_{RR}-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_{SR} & \boldsymbol{G}_{RS}-(\boldsymbol{M}^{-1}N)^{\mathrm{T}}\boldsymbol{G}_{SS}\\ \boldsymbol{N} & M\end{pmatrix}\begin{pmatrix}x_R\\ x_s\end{pmatrix}=\begin{pmatrix}-\boldsymbol{g}_R+(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{g}_s\\ b\end{pmatrix},\tag{12}$$

其中 $\boldsymbol{G}_{RR}$ 是 $\boldsymbol{G}_R$ 取 $j_1,j_2,\cdots,j_p$ 列形成的 $p\times p$ 矩阵,$\boldsymbol{G}_{RS}$ 是 $\boldsymbol{G}_R$ 取 $i_1,i_2,\cdots,i_m$ 列形成的 $p\times m$ 矩阵,$\boldsymbol{G}_{SR}$ 是 $\boldsymbol{G}_S$ 取 $j_1,j_2,\cdots,j_p$ 列形成的 $m\times p$ 矩阵,$\boldsymbol{G}_{SS}$ 是 $\boldsymbol{G}_S$ 取 $i_1,i_2,\cdots,i_m$ 列形成的 $m\times m$ 矩阵。记 $\boldsymbol{C}=\boldsymbol{G}_{RR}-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_{SR}$,$\boldsymbol{D}=\boldsymbol{G}_{RS}-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_{SS}$,$b^*=-\boldsymbol{g}_R+(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{g}_S$,改写方程组(12)得

$$\begin{pmatrix}\boldsymbol{C} & \boldsymbol{D}\\ \boldsymbol{N} & \boldsymbol{M}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}_R\\ \boldsymbol{x}_S\end{pmatrix}=\begin{pmatrix}\boldsymbol{b}^*\\ \boldsymbol{b}\end{pmatrix}。\tag{13}$$

并观察到 $\boldsymbol{M}$ 是非奇异矩阵,用 $\begin{pmatrix}I_p & \boldsymbol{D}\boldsymbol{M}^{-1}\\ 0 & \boldsymbol{M}^{-1}\end{pmatrix}$ 左乘式(13)的两边,得

$$\begin{pmatrix}\boldsymbol{C}-\boldsymbol{D}\boldsymbol{M}^{-1}\boldsymbol{N} & 0\\ \boldsymbol{M}^{-1}\boldsymbol{N} & I_m\end{pmatrix}\begin{pmatrix}\boldsymbol{x}_R\\ \boldsymbol{x}_s\end{pmatrix}=\begin{pmatrix}\boldsymbol{b}^*-\boldsymbol{D}\boldsymbol{M}^{-1}\boldsymbol{b}\\ \boldsymbol{M}^{-1}\boldsymbol{b}\end{pmatrix}。\tag{14}$$

对式(14),若能说明 $\boldsymbol{C}-\boldsymbol{D}\boldsymbol{M}^{-1}\boldsymbol{N}$ 非奇异,则方程组的解易于求得且唯一。

定义 1:对于 $\forall x\in\{x\in R^n\mid \boldsymbol{A}\boldsymbol{x}=0\}$,且 $x\neq 0$,有 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}x>0$,则称 EQP 满足二阶充分条件。

定理 3:若 EQP 满足二阶充分性条件,则 $\boldsymbol{C}-\boldsymbol{D}\boldsymbol{M}^{-1}\boldsymbol{N}$ 非奇异,进而 $\begin{pmatrix}\boldsymbol{G}_S-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_R\\ \boldsymbol{A}\end{pmatrix}$ 也非奇异,方程组(11)的解唯一。

证明:由于 EQP 满足二阶充分性条件,故 $\forall x\in\{\boldsymbol{x}\in R^n\mid \boldsymbol{A}\boldsymbol{x}=0\}$,且 $x\neq 0$,则有 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x}>0$;即当 $(\boldsymbol{N}\ \boldsymbol{M})\begin{pmatrix}\boldsymbol{x}_R\\ \boldsymbol{x}_s\end{pmatrix}=0$ 时,有 $(\boldsymbol{x}_R^{\mathrm{T}}\quad \boldsymbol{x}_S^{\mathrm{T}})\begin{pmatrix}\boldsymbol{G}_{RR} & \boldsymbol{G}_{RS}\\ \boldsymbol{G}_{SR} & \boldsymbol{G}_{SS}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}_R\\ \boldsymbol{x}_S\end{pmatrix}>0$,由 $\boldsymbol{M}$ 非奇异,可得 $\boldsymbol{x}_S=-(\boldsymbol{M}^{-1}\boldsymbol{N})\boldsymbol{x}_R$,有 $(\boldsymbol{x}_R^{\mathrm{T}}\quad \boldsymbol{x}_S^{\mathrm{T}})\begin{pmatrix}\boldsymbol{G}_{RR}\\ \boldsymbol{G}_{RS}\\ \boldsymbol{G}_{SR} & \boldsymbol{G}_{SS}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}_R\\ \boldsymbol{x}_S\end{pmatrix}=\boldsymbol{x}_R^{\mathrm{T}}(I-(\boldsymbol{M}^{-1}\ \boldsymbol{N})^{\mathrm{T}})\begin{pmatrix}\boldsymbol{G}_{RR} & \boldsymbol{G}_{RS}\\ \boldsymbol{G}_{SR} & \boldsymbol{G}_{SS}\end{pmatrix}\begin{pmatrix}I\\ -\boldsymbol{M}^{-1}\boldsymbol{N}\end{pmatrix}\boldsymbol{x}_R=\boldsymbol{x}_R^{\mathrm{T}}(\boldsymbol{G}_{RR}-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_{SR}-(\boldsymbol{G}_{RS}-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_{SS})\boldsymbol{M}^{-1}\boldsymbol{N}))\boldsymbol{x}_R=\boldsymbol{x}_R^{\mathrm{T}}(\boldsymbol{C}-\boldsymbol{D}(\boldsymbol{M}^{-1}\boldsymbol{N}))\boldsymbol{x}_R$,由 $\boldsymbol{x}_R$ 是自由变量,故 $\boldsymbol{C}-\boldsymbol{D}(\boldsymbol{M}^{-1}\boldsymbol{N})$ 正定、非奇异。

结合式(11)–(14)可知 $\begin{pmatrix}\boldsymbol{G}_S-(\boldsymbol{M}^{-1}\boldsymbol{N})^{\mathrm{T}}\boldsymbol{G}_R\\ \boldsymbol{A}\end{pmatrix}$ 也显然非奇异,方程组(11)的解唯一,证毕。

推论 2:若 EQP 满足二阶充分性条件,则 EQP 存在唯一解。

根据上述理论,提出用以获得分类问题最优解的降维二次规划算法。具体算法如下

算法:降维二次规划算法

Step1:用 Gauss 列主元对 $\boldsymbol{A}x=\boldsymbol{b}$ 的增广矩阵进行变换,得到:$\boldsymbol{x}_R+\overline{\boldsymbol{N}}x_S=\overline{\boldsymbol{b}}$,记 $\overline{\boldsymbol{N}}$ 所在列号为 $S=(j_1,j_2,\cdots,j_p)$ 和系数矩阵 $\boldsymbol{A}$ 中其余列号集为 $R=(i_1,i_2,\cdots,i_m)$;

Step2:依据 $\boldsymbol{R}$、$\boldsymbol{S}$ 得到 $\boldsymbol{C}=\boldsymbol{G}_{SR}-\overline{\boldsymbol{N}}^{\mathrm{T}}\boldsymbol{G}_{RR}$,$\boldsymbol{D}=\boldsymbol{G}_{SS}-\overline{\boldsymbol{N}}^{\mathrm{T}}\boldsymbol{G}_{RS}$,$\boldsymbol{b}^*=\overline{\boldsymbol{N}}^{\mathrm{T}}\boldsymbol{g}_R-\boldsymbol{g}_s$;

Step3:$\boldsymbol{x}_s=(\boldsymbol{D}-\boldsymbol{C}\overline{N})^{-1}(\boldsymbol{b}-\boldsymbol{C}\overline{b})$,$\boldsymbol{x}_R=\overline{b}-\overline{N}\boldsymbol{x}_s$,最优解就是 $\begin{pmatrix}\boldsymbol{x}_R\\ \boldsymbol{x}_s\end{pmatrix}$。

2.3　DRPSVM 学习算法

PSVM 的学习过程可以看做是式(3)对应的线性等式约束的二次规划问题,式(3)可转换为矩阵形式

$$\begin{cases}\min & \frac{1}{2}(\boldsymbol{w}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\zeta}^{\mathrm{T}})\boldsymbol{G}(\boldsymbol{w}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}};\\ \text{s.t.} & (\boldsymbol{A}_1,\boldsymbol{A}_2,\boldsymbol{A}_3)(\boldsymbol{w}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}}=\boldsymbol{e},\end{cases}\tag{15}$$

其中 $\boldsymbol{G}=\begin{pmatrix}\boldsymbol{E}_n & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & \boldsymbol{C}\boldsymbol{E}_m\end{pmatrix}$,$\boldsymbol{A}_1=\begin{pmatrix}y_1w_1 & \cdots & y_1w_n\\ \vdots & \ddots & \vdots\\ y_mw_1 & \cdots & y_mw_n\end{pmatrix}$,$\boldsymbol{w}=\begin{pmatrix}w_1\\ \vdots\\ w_n\end{pmatrix}$,$\boldsymbol{\zeta}=\begin{pmatrix}\zeta_1\\ \vdots\\ \zeta_m\end{pmatrix}$,$\boldsymbol{A}_2=$

$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$，$\boldsymbol{A}_3=\boldsymbol{E}_m$，$\boldsymbol{E}_m$ 表示 m 阶单位矩阵，$\boldsymbol{e}$ 表示 $m+n+1$ 维列向量，C 为式(3)的权系数。

记 $\boldsymbol{A}=(\boldsymbol{A}_1,\boldsymbol{A}_2,\boldsymbol{A}_3)$，$\boldsymbol{x}=(\boldsymbol{w}^{\mathrm{T}},\boldsymbol{b}^{\mathrm{T}},\boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}}$，则式(15)可转换为线性等式约束的二次规划问题，故可代入到降维二次规划算法求最优解 x。

传统 PSVM 基于 $K-T$ 条件进行求解，计算复杂度为 $O(n+m)^3$，其中 m 为训练样本个数，n 是训练数据集维数。DRPSVM 的训练计算时间包括降维处理及方程组求解时间，计算复杂度为 $O(n^3+m^3)$，空间复杂度也由 PSVM 所需的 $O(n+m)^2$ 降为 $O(n^2+m^2)$。当 m 和 n 接近时，计算复杂度可降为 PSVM 的 25%左右，而空间复杂度可减少 50%左右。

3 实验与分析

实验是在 CPU 为 Intel P4 3.00 GHZ 的机器上完成的，内存为 1 G，操作系统为 Windows Server 2003。

为了测试 DRPSVM 的推广性能以及分类精度，在 UCI 的 5 个典型的分类数据集上进行实验，实验过程都采用 1/2 的样本作为训练的数据集，另外的 1/2 作为测试数据集，每个算法根据训练数据集的不同选择分别执行 5 次，取它的平均性能作为最后的分类精度。

为了测试分类器对基因芯片数据分类的精度和有效性，选择了来自 CAMDA2000 (international conference for the critical assessment of microarray data analysis)提供的实验数据、文献[23]提供的大肠癌数据一集(colon 1 dataset)和文献[24]提供的大肠癌数据二集(colon 2 dataset)。

CDMDA2000 主要有急性淋巴白血病(ALL)和急性骨髓白血病(AML)2 个基因数据集。在实验设计当中，采用了与文献[25]相同的方法，将该文件数据分为训练集和测试集，分别存储于数据库中，训练集 Trainset 中包含 38 个样本，每个样本包含 7130 个属性，其中最后一个属性为理想的分类结果；测试集 Testset 包含 34 个测试样本，属性为 7129 个。

colon 1 dataset 包含了 62 个样本，每个样本含有 2000 个属性，其中 42 个为大肠癌样本，20 个为正常组织样本，训练集与测试集各占一半；colon 2 dataset 包括 34 个大肠癌样本，其中 20 例为大肠癌原发灶样本，14 例为大肠癌肝转移灶样本，每个样本含有 54675 个属性，训练集与测试集各占一半。

实验采用的 SVM 为 SVMTorch。它是由瑞士 IDIAP 机构开发的。IDIAP 机构是一个半私有且非营利性的研究所，隶属于瑞士联邦科技研究院。SVMTorch 专门为大规模分类问题量身定制，它可以直接处理多种分类问题。它可以从 http://www.idiap.ch/~bengio/projects/SVMTorch.html 上免费获得。

由于 DRPSVM 的时间复杂度和空间复杂度已经从理论上得以证明，实验主要集中在以下几个方面：在 UCI 数据集上与 BP、Nearest、RBF、SVM 分类器进行精度比较；在 CAMDA2000、colon 1 dataset 和 colon 2 dataset 上与 BP、Nearest、RBF、SVM 分类器基因芯片数据分类性能比较。

3.1 在 UCI 数据集上的分类精度比较

表 1 DRPSVM 与其他分类器在 UCI 上的分类精度比较 %

Data set	BP	Nearest	RBF	SVM	DRPSVM
Liver	80.9	69.7	80.3	79.6	75.7
Echo	90.9	90.3	89.4	91.5	90.2
Wine	92.6	84.5	93.6	95.7	94.3
Va-Heart	97.5	95.1	97.0	98.2	97.5
Breast	91.8	92.1	95.2	93.7	93.4

普遍认为，没有哪一种分类算法适合所有的数据集合，同一分类方法在不同的数据集上表现不尽相同。可以看出 DRPSVM 在和 BP、Nearest、RBF、SVM 分类器相比具有次好的分类性能。

从表 1 中可以观察到，在 UCI 的一些数据集上，DRPSVM 较标准 SVM 分类精度稍低，但 DRPSVM 存在唯一最优解，同时训练速度也比标准 SVM 快，因此 DRPSVM 可广泛用于对训练时间敏感的场景，基因芯片数据分析就是这种场景的典型代表。

3.2 在 CAMDA2000 上的基因芯片数据分类性能比较

从实验结果可以明显看出：DRPSVM 和 SVM 相对于 BP、Nearest、RBF 分类器在 CAMDA2000 的 AML 和 ALL 数据集上获得了较好的基因芯片数据分类性能。DRPSVM 在 AML 和 ALL 获得的

平均分类精度是 87.45%，SVM 在 AML 和 ALL 获得的平均分类精度是 87.9%，DRPSVM 还是较 SVM 略低，但是 DRPSVM 的计算复杂度要低，综合评价，DRPSVM 在 CAMDA2000 上具有较好的基因芯片数据分类性能。

表 2　DRPSVM 与其他分类器在 CAMDA2000 上的分类精度比较　%

Data set	BP	Nearest	RBF	SVM	DRPSVM
AML	75.7	63.2	79.6	87.1	86.6
ALL	77.4	67.5	81.4	88.7	88.3

3.3　在 colon 1，2 datasets 上的基因芯片数据分类性能比较

表 3　DRPSVM 与其他分类器在 colon 1，2 datasets 上的分类精度比较　%

Data set	BP	Nearest	RBF	SVM	DRPSVM
colon 1 dataset	83.7	77.6	85.2	93.9	92.8
colon 2 dataset	31.6	32.3	66.8	84.2	85.5

从实验结果可以明显看出：DRPSVM 和 SVM 相对于 BP、Nearest、RBF 分类器在 colon 1 dataset 和 colon 2 dataset 数据集上获得了较好的基因芯片数据分类性能。特别是在 colon 2 dataset 上，当其他分类器由于 colon 2 dataset 的维数急剧升高时，分类性能急剧下滑，但 DRPSVM 在 colon 2 dataset 获得了 85.5%的分类精度，高于 SVM 的 84.2%，原因是 DRPSVM 采用了降维的二次规划算法，特别适用于数据维数高、样本少的应用环境。

4　结　语

笔者提出的 DRPSVM 基因芯片数据分类器继承了 PSVM 的优势，可以将基因芯片数据分类问题转换为仅含线性等式约束的二次规划问题，同时，采用了降维的二次规划算法，使得 DRPSVM 的时间复杂度由传统 PSVM 算法的 $O(n+m)^3$ 降低为 $O(n^3+m^3)$、空间复杂度由 $O(n+m)^2$ 降低为 $O(n^2+m^2)$。该方法与已有主要的分类算法相比，有较好的性能，尽管对于某些基因芯片数据较标准 SVM 算法的精度略低，但训练时间比标准 SVM 算法快，可以满足基因芯片数据分类应用中对训练时间敏感、需要处理小样本、高维数据的苛刻环境，从而具备较大的实用价值。

参考文献：

[1] 顾坚磊，周雁. 中国基因组生物信息学回顾与展望[J]. 中国科学 C 辑：生命科学，2008，38(10)：882-890.
GU JIAN-LEI，ZHOU YAN. Genome bioinformation in China：review and prospective[J]. Science in China Series C：Life Sciences，2008，38(10)：882-890.

[2] SCHENA M，SHALON D，DAVIS R W，et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray[J]. Science，1995，270(5235)：467-470.

[3] HANDL J，KNOWLES J，KELL D B. Computational cluster validation in post-genomic data analysis[J]. Bioinformatics，2005，21(15)：3201-3212.

[4] KHATRI P，DRAGHICI S. Ontological analysis of gene expression data：current tools，limitations，and open problems[J]. Bioinformatics，2005，21(18)：3587-3595.

[5] REIMERS M. Statistical analysis of microarray data[J]. Addiction Biology，2005，10(1)：23-35.

[6] 荆志伟，王忠. 基因芯片数据分析方法研究进展[J]. 生物技术通讯，2007，18(1)：144-148.
JING ZHI-WEI，WANG ZHONG. The methods of classification and analysis of the microarray data[J]. Letters in Biotechnology，2007，18(1)：144-148.

[7] KOOPERBERG C，FAZZIO T G，DELROW J J，et al. Improved background correction for spotted DNA microarrays[J]. Journal of Computational Biology，2002，9(1)：55-66.

[8] DUDOIT R，FRIDLY J，SPEED T P. Comparison of discrimination methods for the classification of tumors using gene expression data[J]. Journal of the American Statistical Association，2002，97(457)：77-87.

[9] HSU A L，TANG S L，HALGAMUGE S K. An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data[J]. Bioinformatics，2003，19(16)：2131-2140.

[10] GUTHKE R，SCHMIDT-HECK W，HAHN D，et al. Gene expression data mining for functional genomics[C]// Proceedings of the European Symposium on Intelligent Techniques，September 14-15，2000，

Aachen, Germany. [S. l.]: IEEE, 2000: 170-177.

[11] ALEXEI A S, DAWOOD B D, MINORU S H K. A web-based tool for principal component and significance analysis of microarray data[J]. Bioinformatics, 2005, 21(10):2548-2549.

[12] CHO J, LEE D, PARK J, et al. Gene selection and classification from microarray data using kernel machine [J]. FEBS Letters, 2004, 571(1-3):93-98.

[13] DANGOND F, HWANG D, CAMELO S, et al. Molecular signature of late-stage human ALS revealed by expression profiling of postmortem spinal cord gray matter[J]. Physiological genomics, 2004, 16(2): 229-239.

[14] FRIEDMAN N, LINIAL M, NACHMAN I, et al. Using bayesian networks to analyze expression data[J]. Journal of Computational Biology, 2000, 7(3-4): 601-620.

[15] IMOTO S, HIGUCHI T, GOTO T, et al. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks[J]. Journal of Bioinformatics and Computational Biology, 2004, 2(1):77-98.

[16] KIM S Y, IMOTO S, MIYANO S. Inferring gene networks from time series microarray data using dynamic bayesian networks [J]. Briefings in Bioinformatics, 2003, 4(3):228-235.

[17] ROSS D T, SCHERF U, EISEN M B, et al. Systematic variation in gene expression patterns in human cancer cell lines[J]. Nature Genetics, 2000, 24(3):227-235.

[18] LIU Y. Active learning with support vector machine applied to gene expression data for cancer classification [J]. Journal of Chemical Information and Computer Sciences, 2004, 44(6):1936-1941.

[19] BROWN M P, GRUNDY W N, LIN D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines[J]. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(1):262-267.

[20] FUREY T S, CRISTIANINI N, DUFFY N. Support vector machine classification and validation of cancer tissue samples using microarray expression data[J]. Bioinformatics, 2000, 16(10):906-914.

[21] WILLIAMS R D, HING S N, GREER B T. Prognostic classification of relapsing favorable histology wilms tumor using cDNA microarray expression profiling and support vector machines[J]. Genes, Chromosomes and Cancer, 2004, 41(1):65-79.

[22] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1-3): 389-422.

[23] ALON U, BARKAI N, NOTTERMAN D A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. Proceedings of the National Academy of the United States of America, 1999, 96(12):6745-6750.

[24] BARRETT T, TROUP D B, WILHITE S E. NCBI GEO: mining tens of millions of expression profiles-database and tools update[J]. Nucleic Acids Research, 2007, 35(S1): 760-765.

[25] GOLUB T R, SLONIM D K, TAMAYO P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999,286(5439):531-537.

(编辑 侯 湘)