

第 2 章 广义的势支持特征选择方法

2.1 引言

特征选择就是从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的，即在一组数量为 D 的特征中选择出数量为 d ($D > d$) 的一组最优特征来，对于分类任务来说，这样做的目的是简化分类算法的时间和空间复杂度，从而提高了分类算法的精度、效率。总的来说，对于分类问题，解决特征选择问题的方法总体上分为两类：一类是将特征选择和分类过程分离，特征选择的过程和具体的分类器无关，如 RELIEF 算法^[13]，这类方法称为过滤型方法 (Filter approach)^[13,52,53]；另外一类是封装型方法 (Wrapper approach)^[54,55]，这类方法将特征选择和分类器的设计融合在一起，特征选择性能和采用的分类器密切相关。两类分法各具特色和不足：滤波型方法实现简单，效率高，但用于分类识别时分类精度较差；封装型方法具有较高的分类性能，但对于不同的分类器，移植性不如滤波法。本章研究的重点属于第二大类，即构造一种封装型的特征选择方法。

近年来，基于支持向量机 SVM 技术的封装型特征选择方法由于其一些独特的优势得到了较为广泛的关注^[2,56,57,58,59,60]。该类方法最突出的优点是基于结构风险最小化原理，具有较强的泛化能力。因此，在特征选择问题上较基于经验风险最小的众多方法具有更好的鲁棒性。

在最近基于支持向量技术的封装型特征选择方法中，S. Hochreiter 和 K. Obermayer 等人提出的势支持向量机 (Potential Support Vector Machine, P-SVM)^[5,61-63] 方法是最具特色的方法之一。该方法通过定义新的目标函数和相应的边界条件直接实现对支持特征的选取，因此和一些基于标准 SVM 的特征选择方法^[57-60] 相比具有以下优点：①由于是直接选取支持特征，而不是通过选择支持向量来实现对特征选取，所以提高了特征选择的速度；②通过定义新的边界条件减少了边缘误差传播的可能性，提高了分类和特征选择的精度和稳定性；③通过定义新的目标函数实现了对非矢量数据集 (Dyadic Matrix) 的处理，从而克服了标准 SVM 方法只能处理成对数据集 (Pairwise Matrix) 的局限，提高了特征选择和分类的适应性。

然而通过研究 Fisher 线性判别准则^[45] 和 P-SVM 对应的目标函数发现，P-SVM 对应的目标函数只是类内离散度的一种特殊情形 (即当各类的样本均值为 $\mathbf{0}$ 时对

应的类内离散度),从而使得该方法的运用受到一定的限制。同时由于满足各类样本均值为 $\mathbf{0}$,一定程度上会导致 $\mathbf{0}$ 矢量周围出现样本交叉,从而不利于P-SVM方法得到最优决策超平面,而且降维后会出现重叠的现象,降低分类效果。针对P-SVM存在的不足,本章通过使用标准的类内散度矩阵重新构造目标函数来实现对P-SVM的泛化,提出一种广义的势支持特征选择方法GPSFM(Generalized Potential Support Features Selection Method)。GPSFM不但提高了P-SVM作为特征选择方法的适应性,而且还继承了势支持向量机作为封装型特征选择方法的优点,同时还具有冗余度底、特征选择快的特点,从而在一定程度上提高了特征选择效果和分类的精度。

2.2 势支持向量机 P-SVM

定义 2.1^[5] 设输入样本空间 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, $\forall \mathbf{x}_i = (x_{i1}, \dots, x_{id})^T$, 复杂特征空间 $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_q\}$, $\forall \mathbf{z}_j = (z_{j1}, \dots, z_{jd})^T$, 则称 $\mathbf{F} = \mathbf{X}^T \mathbf{Z}$ 为并矢矩阵(图 2-1(a)), 显然当 $\mathbf{Z} = \mathbf{X}$ 时, 并矢矩阵转化为成对矩阵(图 2-1(b)), 而当 $\mathbf{Z} = \mathbf{I}$ 时, $\mathbf{F} = \mathbf{X}^T$ 。

	x_1	x_2	x_3	x_4
z_1	0	1	3	-2
z_2	3	-5	1	7
z_3	2	-1	-3	4
z_4	-2	1	-2	0

(a) 并矢数据矩阵

	x_1	x_2	x_3	x_4
x_1	1	-0.1	-2	-3
x_2	-0.1	1	3	1
x_3	-2	3	1	-0.5
x_4	-3	1	-0.5	1

(b) 成对数据矩阵

图 2-1 并矢数据矩阵和成对数据矩阵

从定义 2.1 可以看出, 并矢矩阵 \mathbf{F} 一般是不对称的, 该矩阵的列向量表示输入样本, 行向量代表复杂特征。也就是说, 列向量对应的列对象是要描述的对象, 而复杂特征(行对象)是为描述列对象服务的。比如文献[64,65]中的基因数据集(drug-gene matrix)就是一个并矢矩阵的形式, 该数据集就是使用 drug-cell 矩阵(\mathbf{X} 矩阵)和 gene-cell 矩阵(复杂空间矩阵 \mathbf{Z})相互作用形成的, 因此该数据集使用基因(行对象)来描述生物组织样本(列对象), 而矩阵中任意一个数值则反映了在特定样本中的特定基因的表现程度。再比如在网页文档数据集中, 通常用网页作为输入样本(列对象), 而使用其他包含这些网页(列对象)超链接的网页(行对象)来描述, 数据集中的数值反映了包含直接超链接的数目等。而在实际

的特征选取过程中,对一些特定的数据集所对应的复杂特征空间 Z 不太容易得到,在文献[5,61-63]中 Z 通过两种方法求解:一是假设数据集对应的复杂特征 Z 满足均匀分布或者正态分布,从而随机产生相应的复杂特征空间;二是取 $Z = I$ 。而本章最后测试的数据集就是数据集 X 本身,因此实验过程中设定复杂特征 $Z = I, F = X^T$ 。

图 2-2 表明的是一个二维数据的二分类问题,在图 2-2 (a) 中,两类样本可以较好地被一带有最大间隔的超平面(用实线表示)区分开,其中最大间隔 $\rho = \frac{1}{2}\sqrt{d^2x + d^2y}$, 比率 $\frac{R^2}{\rho^2} = \frac{4R^2}{d^2x + d^2y}$ (该比率与 VC 维相关^[23]);在图 2-2 (b) 中,两类样本(该样本是图 2-2 (a) 中的训练样本沿 y 方向被放大 s 倍后得到的)同样被一带有最大间隔的超平面(实线表示)区分开,其中 $\rho = \frac{1}{2}\sqrt{d^2x + s^2d^2y}$, $\frac{R^2}{\rho^2} = \frac{4R^2}{d^2x + s^2d^2y}$, 而当变换因子 s 再次转换为 $s = 1$ 时,该实线已经不再是原问题的解了(图 2-2 (a) 中的虚线)。因此,在传统的 SVM 方法中,优化超平面在变换因子 s 的影响下并不是不变的(ρ 和 $\frac{R^2}{\rho^2}$ 都随 s 变化而变化),特别是在 SVM 学习方法使用前,训练样本如果已经经过了相应的变换,那么在一定程度上会影响最终的分类效果。

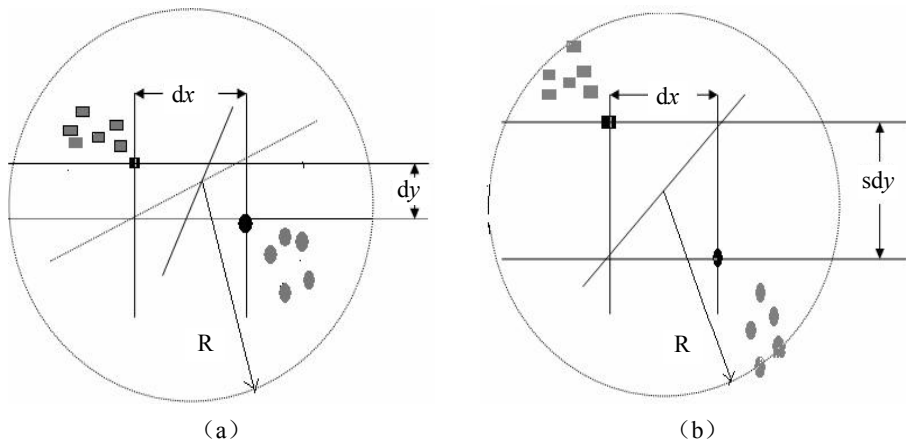


图 2-2 变换因子 s 对选取最优超平面的影响

为了降低变换因子 s 对决策超平面的影响、在一定程度上保持决策超平面稳定性(即不随变换因子 s 的变化而变化),Hochreiter S 和 Obermayer K 等人提出

了势支持向量机 P-SVM。为了便于优化，该支持向量机使用过估计的方法选取比率 $\frac{R^2}{\rho^2}$ 的最小上界 $\|X^T \omega\|$ 作为目标函数，从而在一定程度上保证了决策超平面的不变性，同时根据经验风险最小化和最小二乘原理构造相应的约束条件 $F^T(X^T \omega + b\mathbf{1} - \mathbf{y}) = \mathbf{0}$ ，然而为了降低噪音对特征选择和分类的影响，可以将条件放宽为 $\|F^T(X^T \omega + b\mathbf{1} - \mathbf{y})\|_\infty \leq \varepsilon$ ($\varepsilon > 0$)。该约束条件与传统的 SVM 方法对应的约束条件不同，P-SVM 的每一个约束条件对应的是每一个特征，而传统 SVM 的约束条件对应的是每一个样本，因此相对于直接选取支持向量的传统 SVM，P-SVM 直接选取的是支持特征。由此，可以得到如下 P-SVM 方法的原始优化问题。

定义 2.2^[5] 给定输入样本空间 X 和复杂特征空间 Z 对应的并矢矩阵 F ，输入样本对应的类标号 $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ ，设 ω 、 b 分别为分类超平面对应的法向量、偏移量，则势支持向量机 (P-SVM) 原始最优化问题为：

$$\min_{w,b} \frac{1}{2} \|X^T \omega\|^2 \quad (2.1)$$

$$F^T(X^T \omega - \mathbf{y}) + \varepsilon \mathbf{1} \geq \mathbf{0} \quad (2.2)$$

$$F^T(X^T \omega - \mathbf{y}) - \varepsilon \mathbf{1} \leq \mathbf{0} \quad (2.3)$$

其中 ε 为衡量相关性阈值，决定特征选择的数目。

对定义 2.2 要说明一点，即在 (2.2)、(2.3) 中没有偏移量 b 的出现，这是因为在 P-SVM 方法中已经对矩阵 F 进行了规范化处理^[5,60]，使得 $F^T b \mathbf{1} = \mathbf{0}$ 。

定理 2.1^[5] 考虑线性的 P-SVM 方法，则原始最优化问题 (2.1) ~ (2.3) 的对偶问题是：

$$\begin{aligned} \min_{\alpha^+, \alpha^-} & \frac{1}{2} (\alpha^+ - \alpha^-)^T F^T F (\alpha^+ - \alpha^-) - \mathbf{y}^T F (\alpha^+ - \alpha^-) + \varepsilon \mathbf{1} (\alpha^+ + \alpha^-) \\ & \mathbf{0} \leq \alpha^+ = (\alpha_1^+, \dots, \alpha_0^+)^T \\ & \mathbf{0} \leq \alpha^- = (\alpha_1^-, \dots, \alpha_0^-)^T \end{aligned} \quad (2.4)$$

在势支持向量机方法中，通过定理 2.1 的解实现对支持特征的选取，当 $\alpha_j^+ - \alpha_j^- > 0$ 时，对应的复杂特征 \mathbf{z}_j 被作为支持正类的特征被选取；当 $\alpha_j^+ - \alpha_j^- < 0$ 时，对应的复杂特征 \mathbf{z}_j 被作为支持负类的特征被选取。

定理 2.2^[5] 考虑线性 P-SVM 方法，设 $\alpha = \alpha^+ - \alpha^-$ 是对偶问题 (2.4) 的任意解，则可按下列方式计算出原始问题 (2.1) ~ (2.3) 的解 (ω, b)

$$\omega = Z\alpha \quad (2.5)$$

$$b = \frac{1}{n} \sum_{i=1}^n y_i \quad (2.6)$$

因此，在 P-SVM 算法中，根据定理 2.1 所指定的对偶问题的解来实现特征选择，根据定理 2.2 对应的解生成决策函数 $g(\mathbf{x}_i) = (\boldsymbol{\omega}, \mathbf{x}_i) + b$ 来实现分类。

P-SVM 方法作为一种封装型特征选择方法，既可以实现特征的选取，又可以实现预测样本的分类，而且它是通过求解对偶问题 (2.4) 直接进行特征选择，实验表明有较好的特征选择效果。然而为了便于优化，在寻找比率 $\frac{R^2}{\rho^2}$ 的最小上界 $\|\mathbf{X}^T \boldsymbol{\omega}\|$ 的过程中使用了超估计方法，从而导致 P-SVM 方法存在的问题，这可以使我们通过下文的角度来研究 P-SVM。不难发现，P-SVM 对应的原始问题的目标函数只是各类输入样本均值为 $\mathbf{0}$ 的类内离散度，而在实际问题的求解过程中很少满足该算法所对应的条件，即使满足，也会由于条件的存在导致分类精度在一定程度上受到影响，从而大大地影响了算法的效果和效率。为此，为了在一定程度上提高 P-SVM 方法的适应性和分类效果，本章从更具一般意义的类内离散度出发，泛化 P-SVM 特征选择方法的原始目标函数，提出一种新的广义的势支持特征选择方法——GPSFM。

2.3 广义的势支持特征选择方法：GPSFM

2.3.1 类内离散度

定义 2.3^[45] 假设有 n 个样本组成的样本集 $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ， $\mathbf{x}_i \in \mathbf{R}^d$ ，它们分别属于两个不同的类，其中大小为 n_1 的样本子集 \mathbf{D}_1 属于正类，大小为 n_2 的样本子集 \mathbf{D}_2 属于负类，给定分类决策平面的法向量 $\boldsymbol{\omega}$ ，则类内离散度为 $\boldsymbol{\omega}^T \mathbf{S}_{\boldsymbol{\omega}} \boldsymbol{\omega}$ 。其中：

$$\mathbf{S}_{\boldsymbol{\omega}} = \sum_{k=1}^2 \sum_{\mathbf{x} \in \mathbf{D}_k} (\mathbf{x} - \mathbf{u}_k)(\mathbf{x} - \mathbf{u}_k)^T \quad (\text{称为类内散度矩阵}) \quad (2.7)$$

$$\mathbf{u}_k = \frac{1}{n_k} \sum_{\mathbf{x} \in \mathbf{D}_k} \mathbf{x}, k=1,2 \quad (\text{称为均值}) \quad (2.8)$$

在 Fisher 线性判别准则中，类内离散度是衡量类内样本紧密程度的度量。类内离散度越小，说明类内样本的紧密程度越高，因此，在分类时要尽可能地要求类内离散度小，这样才能保证分类的精度，得到好的分类效果。而提出的 GPSFM 算法就是利用最小化类内离散度作为原问题的目标函数，这样就可以较好地保证分类的效果。

2.3.2 广义的势支持特征选择方法

下面的定理 2.3 和图 2-3 可以帮助我们从另外一个角度来观察 P-SVM。

定理 2.3 在 P-SVM 方法原问题最优化的目标函数是各类样本均值为 $\mathbf{0}$ 的类内离散度。即假设有 m 个样本组成的样本集 $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ ，它们分别属于两个不同的类，其中大小为 m_1 的样本子集 \mathbf{D}_1 属于正类，大小为 m_2 的样本子集 \mathbf{D}_2 属于负类，给定分类决策平面的法向量 $\boldsymbol{\omega}$ ，当样本集均值 $\mathbf{u}_k = \mathbf{0}$ ($k=1,2$ 时)， $\boldsymbol{\omega}^T \mathbf{S}_w \boldsymbol{\omega} = \|\mathbf{X}^T \boldsymbol{\omega}\|^2$ 。

证明 因为

$$\|\mathbf{X}^T \boldsymbol{\omega}\|^2 = \boldsymbol{\omega}^T \mathbf{X} \mathbf{X}^T \boldsymbol{\omega} \quad (2.9)$$

而当 $\mathbf{u}_k = \mathbf{0}$ ($k=1,2$ 时)

$$\boldsymbol{\omega}^T \mathbf{S}_w \boldsymbol{\omega} = \boldsymbol{\omega}^T \sum_{k=1}^2 \sum_{\mathbf{x} \in D_k} \mathbf{x} \mathbf{x}^T \boldsymbol{\omega} \quad (2.10)$$

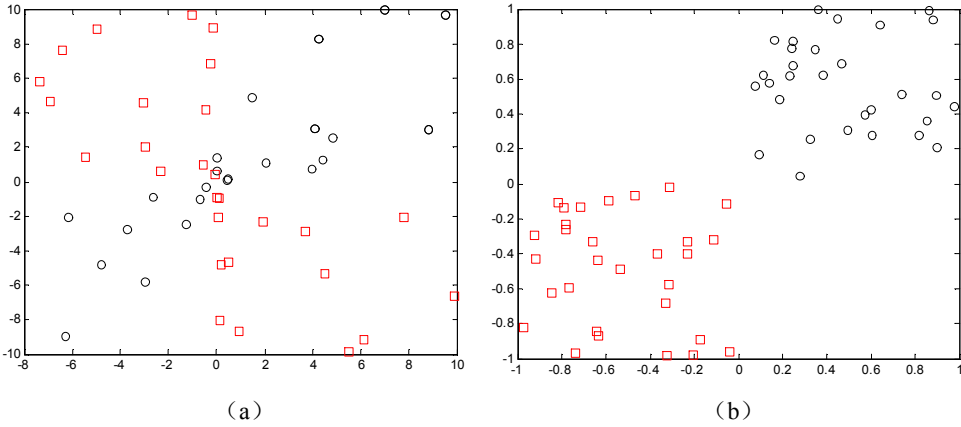
根据基本推导，因 $\mathbf{X} \mathbf{X}^T = \sum_{k=1}^2 \sum_{\mathbf{x} \in D_k} \mathbf{x} \mathbf{x}^T$

所以，定理得证。

从定理 2.3 可以看出，P-SVM 目标函数可以看成当两类样本均值同时为 $\mathbf{0}$ 时的类内离散度，那么在最小化目标函数时，只有保证满足定理条件的特殊的数据集才能达到类内离散度最小。然而在解决实际问题时，绝大部分问题对应的数据集不具备这样的条件，从而使得 P-SVM 这种特征选择算法在处理实际问题时呈现出适应能力不强的特点。

图 2-3 说明的是在二分类时，要求各类均值为 $\mathbf{0}$ 和不要求各类均值为 $\mathbf{0}$ 两种情况可能对分类所产生的影响。从图 2-3 (a) 中可以看出，当要求各类均值为 $\mathbf{0}$ 时，一定程度上会导致在矢量 $\mathbf{0} = (0,0)^T$ 周围出现两类样本相互交叉的情况，从而不利于寻找最优决策超平面，而且样本在降维后在一定程度上会出现重叠的现象，以致降低分类的效果。而当使用一般意义的散度 $\boldsymbol{\omega}^T \mathbf{S}_w \boldsymbol{\omega}$ 来替代 P-SVM 的目标函数时，由于散度矩阵 \mathbf{S}_w 不需要满足各类均值为 $\mathbf{0}$ 的条件，因此一定程度上会防止两类样本交叉重叠的现象出现，有利于寻找最优决策超平面和提高分类精度（图 2-3 (b)）。

而本章广义的势支持特征选择方法 GPSFM 正是将散度 $\boldsymbol{\omega}^T \mathbf{S}_w \boldsymbol{\omega}$ 引入 P-SVM 方法中得到的。因此该方法不但可以提高原方法的适应性，而且还可以在一定程度上提高分类效果和效率，同时具有比较明确的几何含义。



注：红色的“□”、黑色的“○”分别表示不同的两类的样本。

图 2-3 各类样本均值对分类的影响

定义 2.4 给定输入样本空间 \mathbf{X} 和复杂特征空间 \mathbf{Z} 对应的并矢矩阵 \mathbf{F} ，输入样本对应的类标号 $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ ，设 $\boldsymbol{\omega}$ 、 b 分别为分类超平面对应的法向量、偏移量，则 GPSFM 方法的原始最优化问题为：

$$\min_{\boldsymbol{\omega}, b} \frac{1}{2} (\boldsymbol{\omega}^T \mathbf{S}_w \boldsymbol{\omega}) + C \mathbf{1}^T (\boldsymbol{\xi}^+ + \boldsymbol{\xi}^-) \quad (2.11)$$

$$\mathbf{F}^T (\mathbf{X}^T \boldsymbol{\omega} - \mathbf{y}) + \boldsymbol{\xi}^+ + \boldsymbol{\varepsilon} \mathbf{1} \geq \mathbf{0} \quad (2.12)$$

$$\mathbf{F}^T (\mathbf{X}^T \boldsymbol{\omega} - \mathbf{y}) - \boldsymbol{\xi}^- - \boldsymbol{\varepsilon} \mathbf{1} \leq \mathbf{0} \quad (2.13)$$

其中 $\boldsymbol{\xi}^+, \boldsymbol{\xi}^- \geq \mathbf{0}$ 是松弛因子。

松弛因子的作用是放松边界条件，允许有训练集被划错的情况出现。 $\boldsymbol{\varepsilon}$ 为衡量相关性阈值， C 是惩罚参数。阈值 $\boldsymbol{\varepsilon}$ 越大，被选择的特征数越少；而惩罚参数 C 越小，特征选择的冗余度越高^[5]。因此，在实际求解上述对偶式时，要适当地选取这两种参数，以达到较好的特征选择效果。

原始最优化问题目标函数体现了类内离散度最小化，而且实现了对 P-SVM 方法的泛化，同时引入了松弛因子，体现了允许训练集被划错的情况，说明了 GPSFM 具有较强的适应性。下面给出最优化问题 (2.11) ~ (2.13) 的对偶问题。

定理 2.4 考虑线性 GPSFM 方法，则原始最优化问题 (2.11) ~ (2.13) 的对偶问题是：

$$\min_{\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-} \frac{1}{2} (\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-)^T \mathbf{H} (\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) - \mathbf{y}^T \mathbf{F} (\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-) + \boldsymbol{\varepsilon} \mathbf{1} (\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-) \quad (2.14)$$

$$\begin{aligned} & \mathbf{0} < \boldsymbol{\alpha}^+ = (\alpha_1^+, \dots, \alpha_Q^+)^T < C \mathbf{1} \\ \text{s.t.} & \mathbf{0} < \boldsymbol{\alpha}^- = (\alpha_1^-, \dots, \alpha_Q^-)^T < C \mathbf{1} \end{aligned} \quad (2.15)$$

其中： $H = F^T X^T (S_w)^+ X F$ 和 S_w^+ 是 M-P 广义逆，当 S_w 非奇异时，则广义逆为正常逆矩阵，否则为伪逆。

证明 首先对原问题引入 Lagrange 函数。

$$L(\omega, \alpha^+, \alpha^-, \xi^+, \xi^-, \beta^+, \beta^-) = \frac{1}{2}(\omega^T S_w \omega) + C I(\xi^+ + \xi^-) \quad (2.16)$$

$$- \alpha^+ (F^T (X^T \omega - y) + \xi^+ + \varepsilon I) + \alpha^- (F^T (X^T \omega - y) - \xi^- - \varepsilon I) - \beta^+ \xi^+ - \beta^- \xi^-$$

要想原问题达到最优，就必须有

$$\frac{\partial L}{\partial \omega} = 0 \quad (2.17)$$

$$\frac{\partial L}{\partial \xi^+} = 0, \quad \frac{\partial L}{\partial \xi^-} = 0 \quad (2.18)$$

根据 (2.17) 得到

$$\omega = (S_w)^+ X F (\alpha^+ - \alpha^-) \quad (2.19)$$

根据 (2.18) 得到

$$\beta^+ = C I - \alpha^+ \quad (2.20)$$

$$\beta^- = C I - \alpha^- \quad (2.21)$$

将 (2.19) ~ (2.21) 代入 (2.16)，即可得到对偶问题 (2.14) 和 (2.15)。

定理 2.5 考虑线性问题，设 $\alpha = \alpha^+ - \alpha^-$ 是对偶问题 (2.14) ~ (2.15) 的解，则可按下列方式计算出原始最优化问题 (2.11) ~ (2.13) 的解 (ω, b) ：

$$\omega = (S_w)^+ X F (\alpha^+ - \alpha^-) \quad (2.22)$$

$$b = -\frac{1}{m} ((\alpha^+ - \alpha^-)^T F^T X^T (S_w)^+ X - y^T) \mathbf{1} \quad (2.23)$$

证明 (2.22) 已经在定理 2.4 中得到了说明，下面证明 (2.23)。

设在 GPSFM 方法中的决策函数为 $g(x_i) = (\omega, x_i) + b$ ，则对于任意一个样本 x_i ，对应的残差为 $r_i = g(x_i) - y_i$ ，根据最小二乘原理，要想 GPSFM 方法在分类时得到的局部误差最小，就必须满足所有样本对应的残差平方和均值

$R(g(x_i)) = \frac{1}{2n} \sum_{i=1}^n r_i^2$ 最小，即有：

$$\frac{\partial R}{\partial b} = \frac{1}{n} \sum_{x_i} ((\omega, x_i) + b - y_i) = 0 \quad (2.24)$$

因此

$$b = -\frac{1}{n} \sum_{i=1}^n ((\omega, x_i) - y_i) = -\frac{1}{n} (\omega^T X - y^T) \mathbf{1} \quad (2.25)$$

将 (2.22) 代入 (2.25)，即得结果 (2.23)。

在 GPSFM 算法中, 根据定理 2.4 中对偶问题的解来进行支持特征的选取, 当 $\alpha_j^+ - \alpha_j^- > 0$ 时, 对应的复杂特征 \mathbf{z}_j 被作为支持正类的特征被选取; 当 $\alpha_j^+ - \alpha_j^- < 0$ 时, 对应的复杂特征 \mathbf{z}_j 被作为支持负类的特征被选取。然而, 从这种选择过程并不能看出一个具体的复杂特征如何实现对分类的贡献程度以及对偶问题解的直观含义, 因此, 下面给出一个定理来说明。

定理 2.6 考虑线性问题, 设 $\boldsymbol{\alpha}^+ = (\alpha_1^+, \dots, \alpha_Q^+)^T$, $\boldsymbol{\alpha}^- = (\alpha_1^-, \dots, \alpha_Q^-)^T$ 分别是对偶问题 (2.14) 和 (2.15) 的解, 则有下列性质成立:

<1> 当 $\alpha_j^+ > 0$, 则 $\alpha_j^- = 0$;

<2> 当 $\alpha_j^- > 0$, 则 $\alpha_j^+ = 0$;

<3> 对 $\forall j$, $\alpha_j^+ \alpha_j^- = 0$ 。

证明 首先证明性质<1>。

令 $\mathbf{H} = \mathbf{F}^T \mathbf{X}^T (\mathbf{S}_w)^+ \mathbf{X} \mathbf{F}$, $\mathbf{l} = \mathbf{F}^T \mathbf{y}$, $T_j = [\mathbf{H}\boldsymbol{\alpha}]_j - l_j$, 其中 $\boldsymbol{\alpha} = \boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-$, l_j 是 \mathbf{l} 的第 j 个分量, $T_j^+ = T_j + \varepsilon$, $T_j^- = -T_j + \varepsilon$, 因此原始最优化问题的 (2.12) 和 (2.13) 可以表示为:

$$\begin{aligned} \mathbf{F}^T (\mathbf{X}^T \boldsymbol{\omega} - \mathbf{y}) + \boldsymbol{\xi}^+ + \varepsilon \mathbf{1} &= \mathbf{T}^+ + \boldsymbol{\xi}^+ \\ \mathbf{F}^T (\mathbf{X}^T \boldsymbol{\omega} - \mathbf{y}) - \boldsymbol{\xi}^+ - \varepsilon \mathbf{1} &= -(\mathbf{T}^- + \boldsymbol{\xi}^-) \end{aligned} \quad (2.26)$$

根据构造 Karush-Kuhn-Tucker (KKT) 条件的基本原理^[23]、GPSFM 方法的对偶问题和 (2.26), 可以构造如下 KKT 条件:

$$\begin{aligned} \alpha_j^+ (T_j^+ + \xi_j^+) &= 0 \\ -\alpha_j^- (T_j^- + \xi_j^-) &= 0 \\ \beta^+ \xi^+ &= 0 \\ \beta^- \xi^- &= 0 \end{aligned} \quad (2.27)$$

根据前提 $\alpha_j^+ > 0$ 和 KKT 条件 (2.27), 则限制条件 (2.12) 的第 j 个限制条件分量为:

$$[\mathbf{F}^T (\mathbf{X}^T \boldsymbol{\omega} - \mathbf{y})]_j + \xi_j^+ + \varepsilon = T_j^+ + \xi_j^+ = 0 \quad (2.28)$$

因为 $T_j^+ + T_j^- = 2\varepsilon$, 由 (2.28) 可得 $T_j^- = 2\varepsilon + \xi_j^+ > 0$, 所以根据 KKT 条件 (2.27) 可得 $\alpha_j^- = 0$ 。性质成立。

同理可以证明性质<2>。

对于性质<3>, 根据以上两条性质, 当 α_j^+ 和 α_j^- 不同时为 0 时显然成立, 当 α_j^+ 和 α_j^- 同时为 0 时更成立。

从定理 2.6 可以看出, 当 α_j^+ 为非零时, 其对应的复杂特征 \mathbf{z}_j 就是支持正类的

支持特征；当 α_j^- 为非零时，其对应的复杂特征 \mathbf{z}_j 就是支持负类的支持特征。这和标准的 SVM 方法有点相似，只不过这里最后选择的是支持特征，而不是支持向量。而且还可以看出，对于任意一个复杂特征，最多只能对两类中的一类决策起作用，这一点说明了本章的 GPSFM 方法很好地继承了 P-SVM 方法的基本性质，并使该性质在新的泛化后的目标函数上得到了推广。

根据以上定理构造 GPSFM 算法。

算法 2.1 广义的势支撑支持向量机 GPSFM。

Algorithm: GPSFM

Step1: 求解定理 2.4 对应的对偶问题的解: $\boldsymbol{\alpha}^+ = (\alpha_1^+, \dots, \alpha_Q^+)^T$, $\boldsymbol{\alpha}^- = (\alpha_1^-, \dots, \alpha_Q^-)^T$;

Step2: 令 $\alpha_j = \alpha_j^+ - \alpha_j^- (j=1, \dots, Q)$ 。如果 $\alpha_j > 0$ ，则复杂特征空间 \mathbf{Z} 中分量 \mathbf{z}_j 将作为支持正类的特征被选择；反之，将作为支持负类的特征被选择；

Step3: 根据定理 2.5 中的 (2.22)、(2.23)，分别计算分类决策平面法向量 $\boldsymbol{\omega}$ 和偏移量 b ；

Step4: 根据分类决策平面法向量 $\boldsymbol{\omega}$ 和偏移量 b ，构造分类决策函数 $g(\mathbf{x}_i) = (\boldsymbol{\omega}, \mathbf{x}_i) + b$ ；

Step5: 使用决策函数 $g(\mathbf{x}_i)$ 对待分类样本进行分类。

2.4 实验研究

通过以上分析可以看出，GPSFM 方法不但在一定程度上继承了势支持向量机的优点，而且从理论上实现了支持向量机的泛化。下面使用三种线性分类算法（GPSFM、P-SVM、SVM-RFE[58]）分别在国际公认的 UCI 数据集 (<http://www.ics.uci.edu/~mllearn/MLRepository.html>) 和人脸识别数据集 ORL [38,42,68] (<http://www.cs.uiuc.edu/homes/dengcai2/>) 测试，而在测试基因数据集(UCI 中的 DNA 数据集[64,66]、Colon 数据集 (<http://www.molbio.princeton.edu/>) 时只使用 GPSFM 和 P-SVM。需要说明的是，在使用 GPSFM 和 P-SVM 方法时，取复杂特征 $\mathbf{Z} = \mathbf{I}$ ， $\mathbf{F} = \mathbf{X}^T$ ，并对矩阵 $\mathbf{F} = \mathbf{X}^T$ 进行规范化^[5,60]。

在比较 GPSFM 和 P-SVM 方法时，主要说明特征选择结果和测试精度之间的关系，同时说明参数 ε 值对两种方法选择结果的影响。具体实现过程为：在相近的特征选择数目的前提下比较 ε 值和相应的分类精度。而对于 SVM-RFE 方法，采用线性核的方式进行实验。

2.4.1 真实数据

UCI 数据集是经典的数据集，经常用来测试聚类、分类算法的性能。本书使用 UCI 数据集中的多个数据子集来测试 GPSFM 方法，这些数据子集包括：australian 数据集、diabetes 数据集、heart 数据集、breast 数据集、liver_disorder 数据集、ionosphere 数据集、pima 数据集、glass 数据集（float 和 non_float）、sonar 数据集和 threeof9 数据集等，上述数据基本状态见表 2-1。

表 2-1 australian 数据集、heart 数据集数据集

Datasets	The Number of Training Samples	The Number of Test Samples	The Number of Features
australian	460	230	14
diabetes	510	256	8
heart	180	90	13
breast	465	233	10
liver_disorder	230	115	6
ionosphere	230	117	33
pima	510	256	8
glass2	105	54	9
sonar	135	69	60
threeof9	340	170	9

为了说明算法的具体性能，本实验首先对每一个数据集随机组合 5 次，每次组合后采用 5-折交叉验证，并计算得到平均精度和相应的方差。同时为了说明参数 ε 对本书 GPSFM 方法的影响，任意选取 4 个数据集在具体某一次测试过程中的效果来加以说明。实验结果见表 2-2 和图 2-4。

表 2-2 GPSFM 方法、P-SVM 方法和 SVM-RFE 方法测试精度比较

Datasets	The Number of Selected Features	Methods	Accuracies
australian	14	GSPFM	0.73478 ± 0.047029
		P-SVM	0.7058 ± 0.032633
		SVM-RFE	0.72609 ± 0.10237
	12~13	GSPFM	0.76087 ± 0.037148
		P-SVM	0.72457 ± 0.030123
		SVM-RFE	0.80754 ± 0.051871

续表

Datasets	The Number of Selected Features	Methods	Accuracies
australian	10~11	GSPFM	0.77536 ± 0.021447
		P-SVM	0.81043 ± 0.032929
		SVM-RFE	0.79034 ± 0.123
	8~9	GSPFM	0.83072 ± 0.039848
		P-SVM	0.75507 ± 0.048098
		SVM-RFE	0.73623 ± 0.12192
	5~7	GSPFM	0.81884 ± 0.048098
		P-SVM	0.65797 ± 0.031453
		SVM-RFE	0.73118 ± 0.12269
	3~4	GSPFM	0.71426 ± 0.01193
		P-SVM	0.61563 ± 0.026566
		SVM-RFE	0.72309 ± 0.2539
	1~2	GPFM	0.66377 ± 0.018407
		P-SVM	0.60391 ± 0.01368
		SVM-RFE	0.68986 ± 0.14836
diabetes	8	GSPFM	0.66276 ± 0.044244
		P-SVM	0.66797 ± 0.055381
		SVM-RFE	0.47135 ± 0.044936
	6~7	GSPFM	0.73438 ± 0.037263
		P-SVM	0.7487 ± 0.026009
		SVM-RFE	0.60794 ± 0.037292
	3~5	GSPFM	0.76042 ± 0.023868
		P-SVM	0.76832 ± 0.026598
		SVM-RFE	0.52083 ± 0.16798
	1~2	GSPFM	0.65104 ± 0.044936
		P-SVM	0.71224 ± 0.055427
		SVM-RFE	0.42708 ± 0.13104
heart	13	GSPFM	0.81852 ± 0.061195
		P-SVM	0.82963 ± 0.061864
		SVM-RFE	0.76465 ± 0.1753

续表

Datasets	The Number of Selected Features	Methods	Accuracies
heart	11~12	GSPFM	0.82094 ± 0.039021
		P-SVM	0.82593 ± 0.050103
		SVM-RFE	0.74386 ± 0.16814
	8~10	GSPFM	0.84074 ± 0.039021
		P-SVM	0.83333 ± 0.061195
		SVM-RFE	0.79259 ± 0.3407
	6~7	GSPFM	0.84444 ± 0.02222
		P-SVM	0.81481 ± 0.078829
		SVM-RFE	0.82963 ± 0.07229
	4~5	GSPFM	0.8037 ± 0.044905
		P-SVM	0.7963 ± 0.078042
		SVM-RFE	0.6667 ± 0.16814
	2~3	GSPFM	0.68519 ± 0.035717
		P-SVM	0.64676 ± 0.069389
		SVM-RFE	0.6 ± 0.1975
1	GSPFM	0.58779 ± 0.048365	
	P-SVM	0.60741 ± 0.081901	
	SVM-RFE	0.5667 ± 0.18745	
breast	10	GSPFM	0.65522 ± 0.042342
		P-SVM	0.56509 ± 0.01777
		SVM-RFE	0.80963 ± 0.101
	8~9	GSPFM	0.76109 ± 0.084249
		P-SVM	0.70386 ± 0.01777
		SVM-RFE	0.95567 ± 0.016749
	7	GSPFM	0.80687 ± 0.052739
		P-SVM	0.8927 ± 0.060746
		SVM-RFE	0.90844 ± 0.0042918
	5~6	SPFM	0.90275 ± 0.067041
		P-SVM	0.85408 ± 0.060847
		SVM-RFE	0.87139 ± 0.006559

续表

Datasets	The Number of Selected Features	Methods	Accuracies
breast	2~4	GSPFM	0.76402 ± 0.070296
		P-SVM	0.714587 ± 0.060847
		SVM-RFE	0.837842 ± 0.0049558
	1	GSPFM	0.64389 ± 0.052967
		P-SVM	0.55564 ± 0.047281
		SVM-RFE	0.795826 ± 0.0029532
liver_disorder	5~6	GSPFM	0.58841 ± 0.026566
		P-SVM	0.66667 ± 0.04094
		SVM-RFE	0.46667 ± 0.010041
	3~4	GSPFM	0.68261 ± 0.030123
		P-SVM	0.5768 ± 0.027953
		SVM-RFE	0.53623 ± 0.078902
	1~2	GSPFM	0.5791 ± 0.025102
		P-SVM	0.56852 ± 0.023007
		SVM-RFE	0.50725 ± 0.03058
ionosphere	33	GSPFM	0.88889 ± 0.017904
		P-SVM	0.86325 ± 0.037259
		SVM-RFE	0.77778 ± 0.08547
	31~32	GSPFM	0.8661 ± 0.022723
		P-SVM	0.85755 ± 0.044412
		SVM-RFE	0.68946 ± 0.13843
	23~25	GSPFM	0.87179 ± 0.049346
		P-SVM	0.8547 ± 0.043019
		SVM-RFE	0.70573 ± 0.0952
	16~19	GSPFM	0.89744 ± 0.014804
		P-SVM	0.8604 ± 0.013056
		SVM-RFE	0.80912 ± 0.038541
	8~11	GSPFM	0.86352 ± 0.017792
		P-SVM	0.85185 ± 0.071168
		SVM-RFE	0.69231 ± 0.066754

续表

Datasets	The Number of Selected Features	Methods	Accuracies
ionosphere	3~5	GSPFM	0.849 ± 0.021509
		P-SVM	0.8433 ± 0.054281
		SVM-RFE	0.5755 ± 0.09131
	1~2	GSPFM	0.79202 ± 0.065279
		P-SVM	0.71795 ± 0.059215
		SVM-RFE	0.5782 ± 0.1773
pima	8	GSPFM	0.66797 ± 0.024808
		P-SVM	0.75 ± 0.02067
		SVM-RFE	0.46484 ± 0.13003
	6~7	GSPFM	0.75911 ± 0.021514
		P-SVM	0.76823 ± 0.035229
		SVM-RFE	0.47529 ± 0.14894
	3~5	GSPFM	0.76432 ± 0.017027
		P-SVM	0.7526 ± 0.022212
		SVM-RFE	0.48698 ± 0.17834
	1~2	GSPFM	0.66276 ± 0.011934
		P-SVM	0.71094 ± 0.023761
		SVM-RFE	0.42188 ± 0.010335
glass2	8~9	GSPFM	0.69258 ± 0.042767
		P-SVM	0.69136 ± 0.074842
		SVM-RFE	0.67901 ± 0.046604
	5~7	GSPFM	0.66676 ± 0.03529
		P-SVM	0.685391 ± 0.010428
		SVM-RFE	0.67284 ± 0.038549
	3~4	GSPFM	0.59321 ± 0.10692
		P-SVM	0.64815 ± 0.08072
		SVM-RFE	0.62346 ± 0.010692
	1~2	GSPFM	0.48765 ± 0.010593
		P-SVM	0.59295 ± 0.048995
		SVM-RFE	0.5679 ± 0.09135

续表

Datasets	The Number of Selected Features	Methods	Accuracies
sonar	60	GSPFM	0.653892 ± 0.030274
		P-SVM	0.704597 ± 0.083201
		SVM-RFE	0.66547 ± 0.057971
	27~30	GSPFM	0.72446 ± 0.052254
		P-SVM	0.71368 ± 0.066414
		SVM-RFE	0.67633 ± 0.074371
	12~15	GSPFM	0.73745 ± 0.050897
		P-SVM	0.687429 ± 0.047601
		SVM-RFE	0.71498 ± 0.030109
	6~8	GSPFM	0.65739 ± 0.071491
		P-SVM	0.66349 ± 0.052254
		SVM-RFE	0.69565 ± 0.063172
	1~3	GSPFM	0.57971 ± 0.075307
		P-SVM	0.59082 ± 0.11257
		SVM-RFE	0.61353 ± 0.082409
threeof9	9	GSPFM	0.8 ± 0.048149
		P-SVM	0.79412 ± 0.050259
		SVM-RFE	0.78235 ± 0.013362
	7~8	GSPFM	0.7902 ± 0.044927
		P-SVM	0.76471 ± 0.047059
		SVM-RFE	0.73529 ± 0.012959
	3~4	GSPFM	0.59216 ± 0.02227
		P-SVM	0.57059 ± 0.072284
		SVM-RFE	0.58235 ± 0.048149
	1~2	GSPFM	0.56078 ± 0.062898
		P-SVM	0.53333 ± 0.04004
		SVM-RFE	0.56471 ± 0.15478

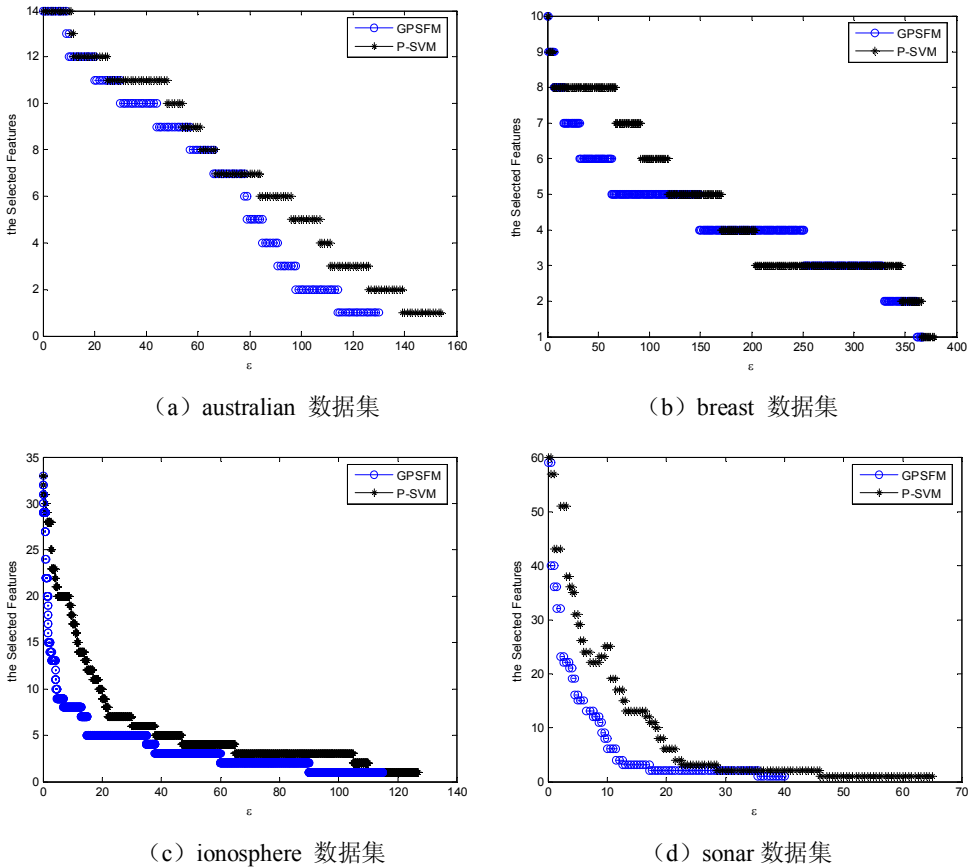


图 2-4 参数与选择特征数的关系

从表 2-2 和图 2-4 可以得到如下结论：

(1) 根据表 2-2 可以看出，在对 UCI 数据集进行测试时，GPSFM 方法只有在 3 个数据集上最高分类精度略低于 P-SVM、SVM-RFE 方法测试所得精度，这可以说明 GPSFM 方法具有较好的分类精度。从测试的方差上看，GPSFM 方法同其他两种方法相比，具有相对较小的方差值，这说明 GPSFM 方法具有较好的分类稳定性。

(2) 从表 2-2 还可以看出，GPSFM 方法符合特征选择方法的一般规律，即测试精度随着特征选择数目的变化有一个从小到大再到小的过程。同时和其他两种方法相比，还具有特征冗余度低的优点，比如：测试 australian 数据集时，GPSFM 方法在支持特征数为 8~9 时测试精度最高，而 P-SVM、SVM-RFE 方法分别在支持特征数为 10~11、12~13 时测试精度最高；在测试 liver_disorder 数据集时，GPSFM 方法在支持特征数为 3 时测试精度最高，而 P-SVM 和 SVM-RFE 则分别

在特征数为 5~6 和 4 时取得最高；在测试 heart 数据集时，GPSFM 方法在支持特征数为 6 时测试精度最高，而 P-SVM、SVM-RFE 方法分别在选择特征数 10、7 左右时测试精度最高。

(3) 图 2-4 说明本章的 GPSFM 方法和 P-SVM 相关性参数 ε 增大，特征选择数减少的基本特点，符合这两种算法构造的基本原理。然而，GPSFM 方法和 P-SVM 相比，当 ε 增大时，拥有更快速的特征选择过程，这就说明 GPSFM 方法具有更强的特征选择能力和效率。

2.4.2 基因数据

基因数据集具有维数高、特征冗余度大的特点，因此经常用来测试分类算法和聚类算法。本章使用一个训练样本和测试样本数较大的基因集（DNA 数据集）和一个维数较高的基因集（Colon 数据集）（表 2-3）来测试 GPSFM 方法的特征选择能力和分类效果。

表 2-3 基因（DNA、Colon）数据集

Datasets	The Number of Training Samples	The Number of Test Samples	The Number of Features
DNA	2000	1186	180
Colon	32	30	2000

表 2-4 GPSFM 方法、P-SVM 方法和 SVM-RFE 方法测试基因数据集精度比较

Datasets	The Number of Selected Features	Methods	Classification Accuracies
DNA	180	GPSFM($\varepsilon=0.1$)	0.66842
		P-SVM($\varepsilon=0.1$)	0.6702
	174~175	GPSFM($\varepsilon=1$)	0.67291
		P-SVM($\varepsilon=1$)	0.6702
	126~128	GPSFM($\varepsilon=10$)	0.6774
		P-SVM($\varepsilon=10$)	0.66896
	57~58	GPSFM($\varepsilon=20$)	0.70257
		P-SVM($\varepsilon=30$)	0.68606
	31~32	GPSFM($\varepsilon=30$)	0.6739
		P-SVM($\varepsilon=40$)	0.67291

续表

Datasets	The Number of Selected Features	Methods	Classification Accuracies
DNA	7	GPSFM($\varepsilon=40$)	0.66994
		P-SVM($\varepsilon=70$)	0.64348
	1	GPSFM($\varepsilon=78$)	0.65896
		P-SVM($\varepsilon=570$)	0.64348
Colon	2000	GPSFM($\varepsilon=0.1$)	0.56552
		P-SVM($\varepsilon=0.1$)	0.5657
	1000~1005	GPSFM($\varepsilon=3$)	0.66897
		P-SVM($\varepsilon=15$)	0.6549
	490~500	GPSFM($\varepsilon=3.42$)	0.71264
		P-SVM($\varepsilon=23$)	0.6827
	250~254	GPSFM($\varepsilon=3.44$)	0.76092
		P-SVM($\varepsilon=23$)	0.6653
	15~18	GPSFM($\varepsilon=3.5$)	0.71264
		P-SVM($\varepsilon=31$)	0.652
	1~2	GPSFM($\varepsilon=8.4$)	0.65289
		P-SVM($\varepsilon=36$)	0.54023

注：以上结果是按表 2-3 比例划分，重复运行 5 次的平均性能

根据表 2-4 可以得到如下结论：

(1) 在测试 DNA 数据集时，尽管测试的样本数较大，然而还可以看出，GPSFM 方法和 P-SVM 方法比较，在特征选择数大致相同的情况下具有较高的分类精度，这说明该方法在处理样本数较高的数据集时具有较好的分类精度和较好的特征选择效果；

(2) 在测试 Colon 数据集时，因为该数据集具有较高的维数，所以从某种程度上更能说明特征选择的效果。从表 2-4 的测试结果来看，GPSFM 方法和 P-SVM 方法相比，不管在特征选择效果还是在分类精度上都具有一定的优势，从而说明了该方法在处理高维基因数据集时可以实现有效的特征选择，以达到提高分类精度的作用；

(3) 根据表 2-4 还可以看出，在测试基因数据集时和测试经典 UCI 数据集时具有相似的结果，即 GPSFM 方法同 P-SVM 方法相比，在特征选择过程中具有更

低的特征冗余度，而且当这两种方法在选择相近的特征数时，GPSFM 方法对应的参数 ε 值要比 P-SVM 方法中相应的参数小得多，这说明本章 GPSFM 方法具有更快速的特征选择过程。

2.4.3 人脸图像数据

ORL (32×32) 人脸识别数据集是经常用来测试特征降维后分类器的分类精度^[38,42,68]，该数据集包含 40 类不同人的脸部图像，每类具有同一个人 10 种不同表情、背景的图像，在本书中，我们选取两类共 20 个人脸图像，其中每类随机选取 3 幅图像作为训练样本，剩下的作为测试样本。图 2-5 表明的是同一个人的 10 幅不同表情的图像，测试结果见表 2-5。



图 2-5 ORL 数据集中某一类所有图像

表 2-5 GPSFM 方法、P-SVM 方法和 SVM-RFE 方法测试人脸数据集精度比较

Datasets	The Number of Selected Features	Algorithms	The Test Set Accuracy
ORL(32×32)	1024	GPSFM($\varepsilon = 0.1$)	0.78571
		P-SVM($\varepsilon = 0.1$)	0.78571
		SVM-RFE	0.85714
	500~512	GPSFM($\varepsilon = 0.3$)	0.85714
		P-SVM($\varepsilon = 4.7$)	0.78571
		SVM-RFE	0.85714
	114~128	GPSFM($\varepsilon = 0.5$)	0.85714
		P-SVM($\varepsilon = 8$)	0.85714
		SVM-RFE	0.92857
	64~88	GPSFM($\varepsilon = 0.85$)	0.92857
		P-SVM($\varepsilon = 11$)	0.71429
		SVM-RFE	0.85714
1~3	GPSFM($\varepsilon = 1$)	0.64286	
	P-SVM($\varepsilon = 19$)	0.42857	
	SVM-RFE	0.5	

从表 2-5 可以看出：

(1) 当 GPSFM 和 SVM-REF 从分类精度上进行比较时，虽然两种方法测试该人脸图像数据时，最高精度都可以达到 0.92875，但 SVM-REF 在选择了 128 个特征后精度达到最高，而本章的 GPSFM 方法在达到最高精度时只选取了 88 个特征；当 GPSFM 同 P-SVM 进行比较时，可以表明 GPSFM 方法具有较高的分类精度，同时具有更低的特征冗余度。

(2) 从 GPSFM 方法、P-SVM 方法、SVM-RFE 方法中选择 1~3 特征进行降维分类的测试结果来看，GPSFM 方法在只选取两个特征时，分类精度还可以达到 0.64286，这充分说明了该方法作为一种封装型特征选择方法，具有一定的特征选择能力和分类效果。

总之，通过以上三个实验可以充分说明 GPSFM 方法具有一定的分类效果，同时表现出较强的特征选择能力，和 P-SVM 相比还具有较快的特征选择速度。

2.5 本章小结

本章通过分析 P-SVM 方法存在的不足，提出一种广义的势特征选择方法 GPSFM，该方法在一定程度上不但继承了原始算法 P-SVM 方法的优点，而且还具有特征选择冗余度低、特征选择速度快、分类精度高、适应性强的特点。当然，在 GPSFM 方法中还存在有待研究的问题，比如参数 ε 和特征选择数之间是否存在一定的线性或非线性的联系，特别是在该方法中引入了类内散度矩阵 \mathbf{S}_w ，从而为构造对应的非线性特征选择方法增加了一定难度，这将是我们的下一步研究的方向。