ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.4, April 2008, pp.842−850
DOI: 10.3724/SP.J.1001.2008.00842

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于分类间隔的特征选择算法*

任双桥+, 傅耀文, 黎 湘, 庄钊文

(国防科学技术大学 电子科学与工程学院空间电子信息技术研究所,湖南 长沙 410073)

Feature Selection Based on Classes Margin

REN Shuang-Qiao+, FU Yao-Wen, LI Xiang, ZHUANG Zhao-Wen

(School of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: Phn: +86-10-66322728, E-mail: shuangqiaoatr@yahoo.com.cn

Ren SQ, Fu YW, Li X, Zhuang ZW. Feature selection based on classes margin. *Journal of Software*, 2008, 19(4):842−850. http://www.jos.org.cn/1000-9825/19/842.htm

Abstract: Firstly, a distinguishable condition is proposed for separating the features by linear classification hyper surface. Secondly, the paper analyses the properties of the feature linear distinguishable criterion based on support vector machines (SVMs). Finally, the efficiency rate of features are defined by the contribution to classes margin of each feature, and a feature selection algorithm is put forward based on the feature efficiency rate. As experimental results show, validated with the actually measuring data and UCI (University of California, Irvine) data, performance of the new feature selection method, such as classification capability and generalized capability are improved obviously in contrast to the classical Relief method.

Key words: feature selection; efficiency rate; classe margin; SVM (support vector machine)

摘 要: 对于二类目标特征选择问题,首先讨论了特征空间的线性可分性问题,并给出了其判别条件;其次,通过借鉴支撑矢量机原理,分析了特征可分性判据的基本性质;最后,依据各特征对分类间隔的贡献大小定义了特征有效率,并以此进行特征选择和特征空间降维.实测数据与网络公开 UCI(University of California,Irvine)数据库的实验结果表明,与经典的 Relief 特征选择算法相比,该算法在识别性能和推广能力上明显有所提高.

关键词: 特征选择;有效率;分类间隔;支撑矢量机

中图法分类号: TP18 **文献标识码**: A

特征选择是模式识别中的关键技术之一[1,2].一般情况下,只有在特征矢量中包含足够多的类别信息,才能通过分类器实现正确分类.由于特征是否包含足够多的信息很难确定,为了提高识别率,总是最大限度地提取特征信息,结果不仅使特征维数增大,而且其中可能存在较多的无效和冗余特征.因此,选择合适的特征来描述模式,对模式识别精度、训练时间和存储空间等许多方面都有较大影响[1],并且对分类器的构造也起着非常重要的作用.

特征选择的标准较多,经典的选择算法大多采用概率度量、距离度量和熵度量等衡量标准[1,3,4].在足够多的样本前提下,基于大样本统计理论的上述准则和相应的分类算法是合理的;而当训练样本为小样本时,上述准则

* Supported by the National Natural Science Foundation of China under Grant No.60402032 (国家自然科学基金)
Received 2006-11-06; Accepted 2007-01-24

不一定总能取得良好的效果[5].此外,特征选择常用的另一类准则还有类内距离、类间距离和散布矩阵度量等[1].这些方法虽然应用广泛,但是,其识别算法大多是建立在经验模型的基础上,如神经网络等[6,7],其模型参数和结构的确定受数据和识别算法的影响较大.总之,经典的特征算法和相应的分类器大多注重的是大样本下使经验风险最小,而对分类器的推广性能要求较低[8−10].

统计学习理论和支撑矢量机(support vector machine,简称SVM)[5]是一种研究有限样本下统计规律及学习方法性质的理论,较好地解决了小样本、非线性、高维数和局部极小点等实际问题.根据SVM分类器及特征选择的特点,近年来,研究人员[11,12]基于SVM将特征选择和分类识别融合在一起,通过利用一定的特征选择标准减少并优化支撑矢量,达到获得最佳特征组合的目的.由于这种方法获得的是直接用于分类的支持矢量子集,因此从理论上看,它将明显优于传统的特征选择方法.文献[11]提出了一种基于余量迭代搜索的多类目标特征选择算法,并讨论了算法的推广性能.Weaton等人[12]基于SVM提出寻找具有最小留一法误差界的特征子集,并通过梯度下降算法代替贪婪算法,分别对模拟数据和真实数据进行了特征选择实验.

对于二类目标识别,本文首先对特征线性可分性给出了判别条件.然后,基于支撑矢量机原理,依据各特征对分类间隔的贡献大小定义了特征有效率,并以此提出了一种新的特征选择算法.该算法秉承了统计学习理论的理论基础——结构风险最小,也即要求选择出的特征子集能够较好地兼顾分类器的分类能力和推广性能.实测数据实验结果表明,与经典的 Relief 特征选择算法相比,本文的算法在识别性能上获得了明显的提高.特别地,对于小样本数据效果更为明显.

1 特征空间目标线性可分条件

1.1 线性可分性定义

对于二类目标识别,给定l个训练样本$D=\{(x_i,y_i),i=1,\ldots,l\}$,其中,$x_i\in R^n$,$y_i\in\{1,-1\}$为类别标示符.如果存在一个线性分类器能将每个样本正确分类,则称样本D线性可分.

定义 1.1. 称样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$,其中,$x_i\in R^n$,$y_i\in\{1,-1\}$线性可分,是指存在$w\in R^n$,$b\in R$,$\rho>0$,使得不等式(1.1)成立:

$$y_i(w^Tx_i+b)\geq\rho,i=1,\ldots,l \tag{1.1}$$

若不存在满足不等式(1.1)的(w,b,ρ),则称样本集 D 线性不可分.如图 1 所示.

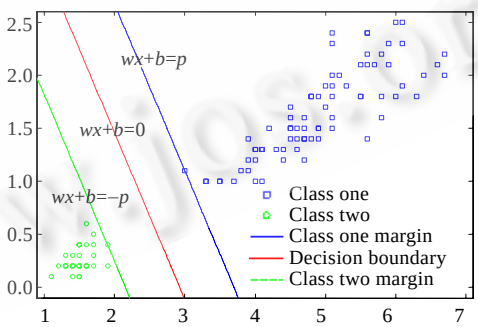

Fig.1 Sample are recognized by linear function

图 1 样本集线性可分示意图

为讨论方便,将不等式(1.1)写成矩阵形式,记$\beta=(w^T,b,\rho)\in R^{n+2}$,$B=(0,\ldots,0,1)^T\in R^{n+2}$,

$$A=\begin{pmatrix}-y_1x_1^T & -y_1 & 1\\ \vdots & \vdots & \vdots\\ -y_lx_l^T & -y_l & 1\end{pmatrix},$$

则有

$$A\beta \le 0 \tag{1.2}$$

$$B^T\beta > 0 \tag{1.3}$$

在下述讨论中,Farkas 引理扮演了一个重要的角色.

引理 1.1(Farkas择一定理)[13]. 设矩阵$A\in R^{m\times n}$,矢量$B\in R^n$,则不等式组①和组②恰一组有解.

组①:$A\beta\le 0, B^T\beta>0$.

组②:$A^T\alpha=B, \alpha\ge 0$.

1.2 线性可分性条件

根据引理 1.1,可从反面考察样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$的线性可分性,也即,若$D$线性不可分,则不等式组①无解,从而其充要条件是不等式组②有解.若将不等式(1.2)、不等式(1.3)代入不等式组②,则可得

$$(y_1x_1\ldots y_lx_l)\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_l\end{pmatrix}=\begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}_{n\times 1} \tag{1.4}$$

$$\sum_{i=1}^{l}\alpha_i=1,\ \sum_{i=1}^{l}\alpha_i y_i=0 \tag{1.5}$$

$$\alpha_i\ge 0, i=1,\ldots,l \tag{1.6}$$

从而可得$D=\{(x_i,y_i),i=1,\ldots,l\}$的线性可分性理论形式的判据:

定理 1.1(可分性条件). 观测样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$线性可分的充要条件是式(1.4)~式(1.6)无解.

2 特征线性可分最优化模型

由定理 1.1 可知,观测样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$线性可分的充要条件是式(1.4)~式(1.6)无解,即下述不等式组无解:

$$H\alpha=0 \tag{2.1}$$

$$\sum_{i=1}^{l}\alpha_i y_i=0,\ \sum_{i=1}^{l}\alpha_i=1 \tag{2.2}$$

$$\alpha_i\ge 0, i=1,\ldots,l \tag{2.3}$$

其中,矩阵$H=(y_1x_1\ldots y_lx_l)$.由于$H\alpha=0$ 的充要条件是$\alpha^TS\alpha=0$,这里,$S=H^TH$, $S_{ij}=y_iy_jx_i^Tx_j$.显然,对称矩阵S半正定,从而对任意的$\alpha\in R^l$,有$\alpha^TS\alpha\ge 0$.因此,可构造如下二次凸规划问题:

$$\begin{aligned}&\min\ \frac{1}{2}\alpha^TS\alpha\\ &\text{s.t.}\quad y^T\alpha=0, e^T\alpha=1\\ &\qquad 0\le\alpha\le 1\end{aligned} \tag{2.4}$$

由定理 1.1 和优化问题式(2.4)可得:

推论 2.1. 观测样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$线性可分的充要条件为二次凸规划问题式(2.4)的目标函数值大于 0.

证明:必要性.假设样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$线性可分,则不等式组(2.1)~不等式组(2.3)无解.因此,对任意满足$y^T\alpha=0$ 和$e^T\alpha=1$ 的非负矢量α,必有$H\alpha\ne 0$,从而可得$\alpha^TS\alpha>0$,故二次凸规划问题式(2.4)的目标函数值大于 0.必要性得证.

充分性.假设二次凸规划问题式(2.4)的目标函数值大于 0,则对任意满足$y^T\alpha=0$ 和$e^T\alpha=1$ 的非负矢量α,必有$\alpha^TS\alpha>0$,从而$H\alpha\ne 0$.因此,不等式组(2.1)~不等式组(2.3)无解,即样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$线性可分.充分性得证.

证毕. □

下面将通过考察支撑矢量机原理[5]来探讨二次凸规划问题式(2.4)的实质.由v-SVM可知,对于样本集D,其最优分类面的求解等价于求解下述二次凸规划问题.

$$\min \frac{1}{2}w^T w-\rho+C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t.}\quad y_i(w^T x_i+b)\geq\rho-\xi_i \tag{2.5}$$
$$\rho\geq 0,\xi_i\geq 0,i=1,\ldots,l$$

其中,w 为特征空间中最优分类超平面的法矢,ξ是引入的松弛变量,C 为正则化参数,$0<C\leq 1$,C 越大,表示对错误分类的惩罚越大.在 v-SVM 模型中,分类面的间隔为 $\frac{\rho}{\|w\|}$.

根据Wolfe对偶理论,二次规化问题不等式(2.5)可以等价于在其对偶空间(即Largrange乘子空间)来求解,其对偶形式为下述二次凸规划问题[5]:

$$\min \frac{1}{2}\alpha^T S\alpha$$
$$\text{s.t.}\quad y^T\alpha=0,e^T\alpha=1 \tag{2.6}$$
$$0\leq\alpha\leq C$$

其中,$0<C\leq 1$,为正则化参数.特别地,当 $C=1$ 时,对偶问题式(2.6)即为凸规划问题式(2.4).因此,满足式(2.1)和式(2.2)的非负权值α,其实质就是 v-SVM 对偶问题式(2.6)中的非负 Largrange 乘子.从而在判断样本集 D 是否线性可分时,同时也实现了支撑矢量机的求解.

3 特征线性可分判据性质

下面来考察当正则化参数$C=1$ 时凸规划问题式(2.6)目标函数值的性质.假设每个观测样本由n个特征构成,即$x_i=(x_{i1},\ldots,x_{in})^T\in R^n,i=1,\ldots,l$,其中,$x_{ij}$为第$j$个特征,则有,

$$S_n=\sum_{k=1}^{n}S_n^k,$$

其中,$S_n^k=(S_{ij}^k)_{l\times l}$,$S_{ij}^k=y_i y_j x_{ik}x_{jk}$,易知 $S_n^k\geq 0$,$k=1,\ldots,n$.令

$$J(n)=\frac{1}{2}\alpha_n^T S_n\alpha_n=\frac{1}{2}\sum_{k=1}^{n}\alpha_n^T S_n^k\alpha_n \tag{3.1}$$

其中,α_n为式(2.6)的解.

定理 3.1. $J(n)$具有单调递增性,即 $J(n)\leq J(n+1)$.

证明:假设在观测样本$x_i=(x_{i1},\ldots,x_{in})^T\in R^n$时,规划问题式(2.6)的解为$\alpha_n$,当样本增加一个特征时,即$x_i=(x_{i1},\ldots,x_{in},x_{in+1})^T\in R^{n+1}$,规划问题式(2.6)的解为$\alpha_{n+1}$,则有,

$$J(n+1)=\frac{1}{2}\alpha_{n+1}^T S_{n+1}\alpha_{n+1}=\frac{1}{2}\sum_{k=1}^{n+1}\alpha_{n+1}^T S_{n+1}^k\alpha_{n+1}$$
$$\geq\frac{1}{2}\sum_{k=1}^{n}\alpha_{n+1}^T S_{n+1}^k\alpha_{n+1}=\frac{1}{2}\sum_{k=1}^{n}\alpha_{n+1}^T S_n^k\alpha_{n+1}$$
$$\geq\frac{1}{2}\sum_{k=1}^{n}\alpha_n^T S_n^k\alpha_n=J(n),$$

其中,第 1 个不等式是由于 $S_{n+1}^k\geq 0$(矩阵 S_{n+1}^k半正定),第 2 个不等式是由于 $J(n)$为 $C=1$ 时规划问题式(2.6)的最优目标函数值.证毕. □

定理 3.2. 假设两个观测样本集$D=\{(u_i,y_i),i=1,\ldots,l\}$,$D_m=\{(v_i,y_i),i=1,\ldots,l\}$,其中,$u_i\in R^n$,$v_i\in R^m$.令样本集$D_{m+n}=\{(x_i,y_i),i=1,\ldots,l\}$,$x_i=(u_i,v_i)\in R^{n+m}$,则有

$$J(n)+J(m)\leq J(n+m),$$

其中,$J(n),J(m),J(n+m)$分别为样本集D_n,D_m,D_{n+m}所对应的规划问题式(2.6)在正则化参数$C=1$ 时的最优目标函数值.

证明:由于$x_i=(u_i,v_i)\in R^{n+m}$,则有

$$S_{n+m}=R_{n+m}^TR_{n+m}=\begin{pmatrix} x_1^Tx_1 & \cdots & y_1y_lx_1^Tx_l \\ \vdots & \vdots & \vdots \\ y_1y_lx_1^Tx_l & \cdots & x_l^Tx_l \end{pmatrix}$$

$$=\begin{pmatrix} u_1^Tu_1+v_1^Tv_1 & \cdots & y_1y_l(u_1^Tu_l+v_1^Tv_l) \\ \vdots & \vdots & \vdots \\ y_1y_l(u_1^Tu_l+v_1^Tv_l) & \cdots & u_l^Tu_l+v_l^Tv_l \end{pmatrix}$$

$$=S_n+S_m,$$

从而有,

$$J(n+m)=\frac{1}{2}\alpha_{n+m}^TS_{n+m}\alpha_{n+m}$$

$$=\frac{1}{2}\alpha_{n+m}^TS_n\alpha_{n+m}+\frac{1}{2}\alpha_{n+m}^TS_m\alpha_{n+m}$$

$$\geq\frac{1}{2}\alpha_n^TS_n\alpha_n+\frac{1}{2}\alpha_m^TS_m\alpha_m$$

$$=J(n)+J(m),$$

其中,$\alpha_n,\alpha_m,\alpha_{n+m}$分别为样本集$D_n,D_m,D_{n+m}$所对应的规划问题式(2.6)在正则化参数$C$=1 时的最优解. □

定理 3.3. 假设两个观测样本集$D_n=\{(u_i,y_i),i=1,\ldots,l\}$,$D_m=\{(v_i,y_i),i=1,\ldots,l\}$,其中,$u_i\in R^n$,$v_i\in R^m$.若$J(n)\leq J(m)$,则有 $\Delta_n^*\leq\Delta_m^*$,其中,$J(n)$,$J(m)$分别为样本集D_n,D_m所对应的规划问题式(2.6)在参数C=1 时的最优目标函数值. $\Delta_n^*=\frac{\rho_n^*}{\|w_n^*\|}$,$\Delta_m^*=\frac{\rho_m^*}{\|w_m^*\|}$ 分别为样本集D_n和D_m的最优分类间隔.

证明:不妨假设$J(n)>0$,若不然,由推论 2.1 可知,结论显然成立.因此,样本集D_n和D_m均线性可分.由支撑矢量机的原理可知,线性可分样本集$D_n=\{(u_i,y_i),i=1,\ldots,l\}$的最优分类间隔可由如下凸规划问题求得,也即

$$\min\frac{1}{2}w^Tw-\rho$$

$$\text{s.t.}\quad y_i(w^Tu_i+b)\geq\rho \tag{3.2}$$

$$\rho\geq0,i=1,\ldots,l$$

由 Wolfe 对偶原理可知,凸规划问题不等式(3.2)的对偶形式如下:

$$\max-\frac{1}{2}\alpha^TS_n\alpha$$

$$\text{s.t.}\quad y^T\alpha=0,e^T\alpha=1 \tag{3.3}$$

$$0\leq\alpha\leq1$$

其中,α为非负Largrange乘子, $w=\sum_{i=1}^{l}\alpha_iy_iu_i$. 设$\alpha_n$为凸规划问题式(3.3)的最优解,则有,

$$J(n)=\frac{1}{2}\alpha_n^TS_n\alpha_n=\frac{1}{2}\|w_n^*\|^2 \tag{3.4}$$

$$\frac{1}{2}\|w_n^*\|^2-\rho_n^*=-\frac{1}{2}\|w_n^*\|^2 \tag{3.5}$$

由式(3.4)、式(3.5)可得:

$$\rho_n^*=\|w_n^*\|^2 \tag{3.6}$$

从而,样本集D_n的最优分类间隔为

$$\Delta_n^*=\sqrt{2J(n)} \tag{3.7}$$

同理,线性可分样本集$D_m=\{(v_i,y_i),i=1,\ldots,l\}$的最优分类间隔为

$$\Delta_m^*=\sqrt{2J(m)} \tag{3.8}$$

由于 $J(n)\le J(m)$,因此 $\Delta_n^*\le\Delta_m^*$. □

定理 3.4. 若样本集$D=\{(u_i,y_i),i=1,\ldots,l\}$线性不可分,且设$u_i=(u_{i1},\ldots,u_{in})^T\in R^n$,$u_{ij}$为第$j$个特征,则任意剔除$m>0$个特征后组成新的样本集仍然线性不可分.若样本集$D$线性可分,则增加任意$m>0$ 个特征后组成新的样本集仍然线性可分,且其分类间隔不会下降.

证明:假设样本集$D=\{(u_i,y_i),i=1,\ldots,l\}$线性不可分,则由推论 2.1 可知,当正则化参数$C=1$ 时,有

$$J(n)=\frac{1}{2}\alpha_n^T S_n\alpha_n=0,$$

则任意剔除 $m>0$ 个特征后,由定理 3.1 可知:

$$0\le J(n-m)\le J(n)=0.$$

从而,$J(n-m)=0$,即新的样本集线性不可分.同理,假设样本集$D=\{(u_i,y_i),i=1,\ldots,l\}$线性可分,则有$J(n)>0$,增加任意$m>0$ 个新特征后,则有 $0<J(n)\le J(n+m)$,也即新的样本集线性可分.

当新增$m>0$ 个特征后,假设样本集$D_{m+n}=\{(x_i,y_i),i=1,\ldots,l\}$,$x_i=(u_i,v_i)\in R^{n+m}$,则由$J(n)\le J(n+m)$和定理 3.3 可知,$\Delta_n^*\le\Delta_{n+m}^*$.因此,当样本新增 $m>0$ 个特征后,其最优分类间隔随特征维数的增加而递增.证毕. □

综上所述,对于观测样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$,当正则化参数$C=1$ 时,二次凸规划问题式(2.6)的目标函数值$J(n)$具备了作为类别可分性判据的下述基本条件[1]:

① $J(n)$与误判概率的界存在单调关系.

根据统计学习理论,期望风险 $R(\alpha)=\int Q(z,\alpha)\mathrm{d}F(z)$ 与经验风险 $R_{emp}(\alpha)=\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)$,至少以 $1-\eta$的概率满足下述不等式:

$$R(\alpha)\le R_{emp}(\alpha)+\Phi(h,l,\eta),$$

其中,$Q(z,\theta),\alpha\in\Omega$为决策风险函数,$z_i=(x_i,y_i)$为观测样本.$\Phi$为致信范围,$h$为函数集$Q(z,\alpha),\alpha\in\Omega$的VC维,且$\Phi(h,l,\eta)$是关于VC维$h$的单调增函数,而关于样本规模$l$是单调减函数.

在统计学习理论中,关于线性函数集的 VC 维有一个重要结论.

定理 3.5[5]. 设矢量$x\in R^n$包含在一个半径为r的超球中,则分类间隔为Δ的超平面集合的VC维h以下述不等式为界:

$$h\le\min\left(\left[\frac{r^2}{\Delta^2}\right] \quad n\right)+1.$$

由此可得关于测试样本误判概率的一个上界,即测试样本不能被分类间隔为Δ的超平面正确分类的概率P_{error}至少以 $1-\eta$的概率满足下述不等式:

$$P_{error}\le\frac{l_0}{l}+\frac{\varepsilon}{2}\left(1+\sqrt{1+\frac{4l_0}{l\varepsilon}}\right),$$

其中,

$$\varepsilon=4\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{4}}{l}.$$

l_0是训练样本中没有被分类间隔为Δ的超平面正确分类的样本数.

因此,由定理 3.5 可知,超平面函数集的VC维h是分类间隔Δ的单调减函数.从而有,当采用超平面进行分类识别时,期望风险$R(\alpha)$的界和测试样本误判概率P_{error}的界均为分类间隔的单调减函数.又由于最优分类间隔$\Delta_n^*=\sqrt{2J(n)}$,从而有,期望风险$R(\alpha)$和测试样本误判概率P_{error}的界均为$J(n)$的单调减函数.

② $J(n)$具有不完全可加性.

③ 由于 $J(n)=\frac{1}{2}\|w_n^*\|^2$,因此,$J(n)$具有“距离”的某些共性.

④ $J(n)$关于特征维数是单调递增的.

4 基于分类间隔的特征选择

由于$J(n)$具备以上 4 个基本条件,因此,可以将其作为类别可分性判据[1].在此基础上,可定义各特征的有效率并以此进行特征选择.假设样本集$D=\{(x_i,y_i),i=1,\ldots,l\}$线性可分,则 $J(n)=\frac{1}{2}\alpha_n^T S_n\alpha_n=\frac{1}{2}\sum_{k=1}^{n}\alpha_n^T S_n^k\alpha_n>0$,从而,可以定义第 k 个特征的有效率为

$$\gamma_k=\frac{\alpha_n^T S_n^k\alpha_n}{\alpha_n^T S_n\alpha_n},k=1,\ldots,n.$$

于是,

$$\sum_{k=1}^{n}\gamma_k=1,\gamma_k\geq 0.$$

由于 $w_n^*=\sum_{i=1}^{l}\alpha_i^n y_i x_i$,$\|w_n^*\|^2=\alpha_n^T S_n\alpha_n$,且 $S_n^k=(S_{ij}^k)_{l\times l}$,$S_{ij}^k=y_i y_j x_{ik}x_{jk}$. 因此,

$$\gamma_k=\left(\frac{w_n^*(k)}{\|w_n^*\|}\right)^2,k=1,\ldots,n \tag{4.1}$$

其中,$w_n^*=(w_n^*(1),\ldots,w_n^*(n))^T$.由式(4.1)可知,第$k$个特征的有效率$\gamma_k$的实质就是最优分类超平面法矢 w_n^* 的第k个分量 $w_n^*(k)$ 占其总能量$\|w_n^*\|^2$ 的百分比.又由于最优分类间隔 $\Delta_n^*=\sqrt{2J(n)}=\|w_n^*\|$,因此,$\gamma_k$很好地刻画了第$k$个特征对分类间隔 Δ_n^* 的贡献.

下面考察在一些特殊情形下特征的有效率,以验证式(4.1)可以很好地吻合直观感知.假设第k_1个特征和第k_2个特征满足如下线性关系:

$$x_{ik_1}=\theta x_{ik_2}+\sigma,i=1,\ldots,l \tag{4.2}$$

不妨假设,$|x_{ik_1}|\leq|x_{ik_2}|$,也即 $0<|\theta|\leq 1$,则有,

$$w_n^*(k_1)=\sum_{i=1}^{l}\alpha_i^n y_i x_{ik_1}=\theta\sum_{i=1}^{l}\alpha_i^n y_i x_{ik_2}=\theta w_n^*(k_2) \tag{4.3}$$

从而可得:

$$\gamma_{k_1}=\theta^2\gamma_{k_2} \tag{4.4}$$

因此,由式(4.4)可知,当第k_1个特征和第k_2个特征完全线性相关时,其特征的有效率是以θ^2下降的.由于大多数分类器均采用距离度量,因此,数值较小的特征提供的有效信息应不如数值较大的特征.特别地,当$\theta=0$ 时,即 x_{ik_1} 恒等于常数σ,则其有效率 $\gamma_{k_1}=0$,这一结论也符合客观事实.因此,式(4.1)的特征有效率与直观感知的吻合度是很好的.

假设二次凸规化问题式(2.6)的目标函数值 $J(n)=0$,也即观测样本集 D 线性不可分.此时,由线性不可分情形下 v-SVM 原理可知,只需对 Langrange 乘子α的边界约束条件作进一步限制,即要求 $0\leq\alpha\leq C$.这里,C 为正则化参数,可设定 $0<C<1$,也即在分类间隔和误判概率之间作一个折衷以确保 $J(n)>0$.总之,无论观测样本集 D 是否线性可分,均可根据特征有效率的大小进行特征的剔除和增加.

5 实验结果及分析

本节利用实测数据和公开的UCI(University of California,Irvine)数据库进行实验,并与经典的Relief特征选择算法进行了比较,取得了较好的结果.Relief[7,14]是Kira等人 1992 年提出的公认的性能较好的特征选择方法,该算法认为,一个好的特征应该使最近邻的同类样本之间特征值相同或相近,而使最近邻的不同类样本之间的值不同或者差别很大.因此,若赋予每个特征相应权值并进行特征排序,通过设定特征权阈值或者特征子集的数目即可进行相应的特征选择.

本次实验数据为某机载雷达实测一维距离像数据集,共有 3 939 个样本,样本维数为 32,包括了 15 个角度下的测量值,样本类型包括卡车和坦克两种目标.该数据集已经过距离对准和能量归一化等预处理过程,且为了更好地对比,本实验将选择一个小样本训练集.实验中选取的初始训练样本集为每个角度下若干个样本,总共选取了 100 个训练样本,测试样本为 3 839 个,依据特征有效率γ大小选择的特征维数从 4 开始依次递增,采用 SVM 方法进行测试识别,实验结果如图 2(a)所示.

第 2 组实验数据为网络上公开的 UCI 数据库中的某癌症患者数据集,共有 569 个样本,样本维数为 30,样本类型包括健康者和患者两种类型.实验中选取的初始训练样本集为 100 个,测试样本为 469 个,依据特征有效率γ大小选择的特征维数从 14 开始依次递增,采用 SVM 方法进行测试识别,实验结果如图 2(b)所示.

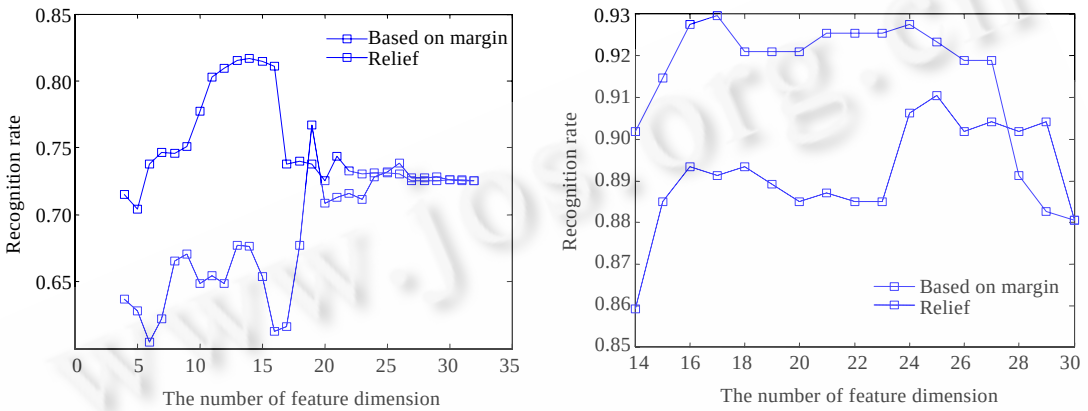

Fig.2 Relation between number of feature dimension and recognition rate

图 2 特征维数与测试样本分类正确率之间的关系

由图 2(a)可知,当特征维数等于 14 时,本文算法的测试分类结果最好,测试率为 81.69%,选出的特征为[1 12 15 16 17 18 19 20 22 23 26 27 29 32].当全部特征都选用时,测试率为 72.57%,相对于最优特征集的测试率下降了 9.12%.而由 Relief 算法选出的最优特征集为 19 个,测试率为 76.69%,低于本文算法的最优测试率.

由图 2(b)可知,当选择特征维数等于 17 时,测试分类结果最好,测试率为 92.96%,选出的特征子集为[1 5 6 7 8 9 11 12 13 16 17 19 25 26 28 29 30].当全部特征都选用时,测试率为 88.06%,相对于最优特征集的测试率下降了 4.9%.由 Relief 算法选出的最优特征集为 25 个,测试率为 91.04%,低于本文算法的最优测试率.

6 结束语

对于二类目标识别问题,本文首先针对样本空间中的可分性给出了其判别条件;然后,基于 SVM 原理探讨了特征选择问题,以特征有效率为切点,提出了一种新的特征选择算法;最后,通过两个小样本实测数据实验验证了该特征选择算法较之经典的 Relief 算法在识别性能和推广能力上更为有效.

References:

[1] Sun JX, *et al*. Modern Pattern Recognition. Changsha: National University of Defense Technology Press, 2002 (in Chinese).

[2] Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. on Knowledge and Data Engineering, 2005,17(4):491−502.

[3] Fan JS, Fang TJ. Analysis and evaluation on main factors for feature selection and abstraction. Computer Engineering and Appliction, 2001,13(2):95−99 (in Chinese with English abstract).

[4] Kwak N, Choi CH. Input feature selection for classification problems. IEEE Trans. on Neural Networks, 2002,13(1):143−159.

[5] Vapnik VN. Statistical Learning Theory. Beijing: Publishing House of Electronics Industry, 2004.

[6] Garrett D, Peterson DA, Anderson CW, Thaut MH. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. IEEE Trans. on Neural Systems and Rehabilitation Engineering, 2003,11(2):141−144.

[7] Zhang LX, Wang JX, Zhao YN, Yang ZH. Combination feature selection based on relief. Journal of Fudan University (Natural Science), 2004,43(5):893−898 (in Chinese with English abstract).

[8] Jin X, Deng YF, Zhong YX. Mixture feature selection strategy applied in cancer classification from gene expression. In: Proc. of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conf. Shanghai: IEEE Press, 2005. 4807−4809.

[9] Xu JQ, Yuan ZD. A feature selection method based on minimizing generalization bounds of SVM Via GA. In: Proc. of the 2003 IEEE Int'l Symp. on Intelligent Control Houston. Texas: IEEE Press, 2003. 996−999.

[10] Sindhwani V, Rakshit S, Deodhare D, Erdogmus D, Principe JC, Niyogi P. Feature selection in MLPs and SVMs based on maximum output information. IEEE Trans. on Neural Networks, 2004,15(4):937−948.

[11] Gilad-Bachrach R, Navot A, Tishby N. Margin based feature selection—Theory and algorithms. In: Proc. of the 21st Int'l Conf. on Machine Learning. Banff: ACM Press, 2004. 43−50.

[12] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. In: Advances in Neural Information Processing Systems, Vol.13. Cambridge: MIT Press, 2000. 668−674.

[13] Su TS, *et al*. Optimization Computation Principle and Program Designing. Changsha: National University of Defense Technology Press, 2001 (in Chinese).

[14] Li Y, Wu ZF, Liu JM, Tang YY. Efficient feature selection for high-dimensional data using two-level filter. In: Proc. of the 3rd Int'l Conf. on Machine Learning and Cybernetics. Shanghai: IEEE Press, 2004. 1711−1716.

附中文参考文献:

[1] 孙即祥,等.现代模式识别.长沙:国防科技大学出版社,2002.

[3] 范劲松,方廷健.特征选择和提取要素的分析及其评价.计算机工程与应用,2001,13(2):95−99.

[7] 张丽新,王家,赵雁南,杨泽红.基于 Relief 的组合式特征选择.复旦大学学报(自然科学版),2004,43(5):893−898.

[13] 粟塔山,等.最优化计算原理与算法程序设计.长沙:国防科技大学出版社,2001.

任双桥(1977—),男,湖南望城人,博士,主要研究领域为支撑矢量机,目标识别.

黎湘(1967—),男,博士,教授,主要研究领域为目标识别,信息融合.

傅耀文(1976—),男,博士,副教授,主要研究领域为目标识别,信息融合.

庄钊文(1958—),男,博士,教授,CCF 高级会员,主要研究领域为信息融合,雷达极化信息处理.