# 基于先验知识的多类CVM航班延误预警模型

袁 瑗[1],陈 兵[1],徐 涛[2],王建东[1]

(1.南京航空航天大学 信息科学与技术学院,南京 210016;2.中国民航大学 计算机学院,天津 300000)

**摘 要**:基于使用现有的支持向量机解决机场航班延误预警问题存在未充分利用先验知识和训练需花费大量时间和空间的问题,提出了基于中心约束最小闭包球的加权多类算法。该算法首先利用先验知识确定一种新的基于相对紧密度的方法计算样本权值并将其融合到支持向量机中,然后转化为中心约束的最小闭包球进行训练。实验结果表明,该方法比现有的支持向量机具有更合理的分类面并且训练速度得到大大提高。

**关键词**:人工智能;航班延误;支持向量机;最小闭包球;先验知识

**中图分类号**:TP18 **文献标志码**:A **文章编号**:1671-5497(2010)03-0752-06

# Prior knowledge based multi-class core vector machine for flight delay early warning

YUAN Yuan[1], CHEN Bing[1], XU Tao[2], WANG Jian-dong[1]

(1. *College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*; 2. *College of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300000, *China*)

**Abstract**: The early warning of airport runtime flight delay is a multi-class classification problem. There are two issues when solving this problem using the normal Support Vector Machine (SVM). The first issue is that the prior knowledge is not adequately utilized, and the second issue is intensive time and space consumption for data training. A new algorithm, which is called as center-constrained Minimum Enclosing Ball (MEB) based weighted margin multi-class algorithm is proposed. First, the proposed algorithm uses the prior knowledge to build a new methodology which is based on a new relative affinity function. Then this new methodology is used to calculate the weights of the sample data and add them to the SVM. After adding these features, the SVM is converted to a center-constrained MEB and can be trained easily. Experiments show that the proposed algorithm not only gives more reasonable classification results comparing to normal SVM, but also obviously speeds up the data training processing.

**Key words**: artificial intelligence; fight delay; support vector machine; minimum enclosing ball; prior knowledge

**收稿日期**:2008-08-03.

**基金项目**:"863"国家高技术研究发展计划项目(2006AA12A106).

**作者简介**:袁瑗(1984-),女,博士研究生.研究方向:数据挖掘. E-mail:yuanyuanlucky2006@yahoo.com.cn

**通信作者**:徐涛(1962-),男,教授.研究方向:数据挖掘,民航信息系统理论,图形图像与可视化技术.
E-mail:taoxucs@nauu.edu.cn

利用预警管理理论对机场航班进行延误等级的预警将是一种有效缓解航班延误损失的手段。由 Vapnik 等[1]提出的 SVM(支持向量机理论)是处理非线性问题的有效方法。SVM 的特点启示我们可以用该方法建立航班延误预警模型。与大多数机器学习方法一样，SVM 是一种基于统计的纯数据驱动方法，需要足够多标记好的训练样本才能保证模型的性能，而这个条件在应用中往往不可得。若能获得先验知识并将其融合到 SVM 中，就可以弥补标签样本不足的缺陷[2-3]。Schölkopf 等[4]给出了通过核函数结合先验知识的方法。2001 年，Fung[5]给出了通过多面体集重构 SVM，从而引入先验知识的方法。2004 年，Wu[2]在给定数据集属于某类的置信度的基础上，提出了可以融合先验知识的加权间隔(weighted margined svm，WMSVM)，该方法可以获得更合理的分类面。

Wu[2]方法存在两个局限：①在使用通常的 SVM 训练时，时间复杂度为 $o(m)$ 到 $o(m^{2.3})$ [6]。虽然可以用并行技术提高到 $o(m)$ [7]，但并没有理论保证。2005 年 Tsang 等[8]提出了基于最小闭包球的支持向量机算法 CVM(Core vector machine)，它的核心思想是将支持向量机的二次规划问题转化为 MEB(最小闭包球)问题，然后利用有效的 $(1+\varepsilon)$ 近似算法来求解。理论证明该方法的时间复杂度与样本大小成线性关系，空间复杂度与样本无关，有效地解决了大样本的训练问题。在 2005 年 Tsang 等[9]进一步在 CVM 算法中采用 CMEB(中心约束最小闭包球，Center-constrained MEB，CMEB)替代 MEB，从而打破了核函数必须为同一常数的限制；②Wu 方法只讨论了两类问题。Asharaf[10]提出了多类 CVM (Multiclass core vector machine)算法，很好地解决了多类问题。本文将先验知识融合到多类 CVM，并转化为中心约束 MEB 来训练。试验表明，基于融合先验知识的多类 CVM 航班延误预警模型比多类 CVM 算法有更好的分类效果，比 Wu[2]的方法有更快的训练速度。

## 1　问题的提出

当机场容量不能满足进离港航班需求量时将产生航班延误，由于航班计划通常不是均匀分布的，因而单位间隔的进离港航班需求量是变化的；而机场容量因为受跑道、天气、进离港航班比率等因素影响也并非确定不变。图 1 描述了首都机场某日不同时段的计划进离港航班需求量和机场容量，纵轴表示航班架次，横轴表示时间，取样间隔为 30 min。图 2 为航班延误数量与进离港航班需求量之间的非线性关系：当需求量远小于机场容量时，航班延误的变化率较小；当需求量接近或大于机场容量时，航班延误的变化率较大。事实上，当跑道在重负荷下使用时，任何间隔内都表现出一种复杂的动态随机行为[11]。

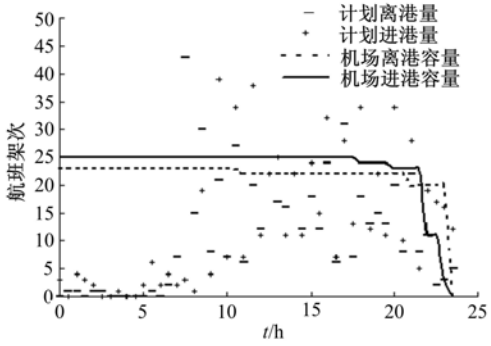


**图 1　进离港航班需求量与容量图**

**Fig. 1　The requirement and capacity curve**

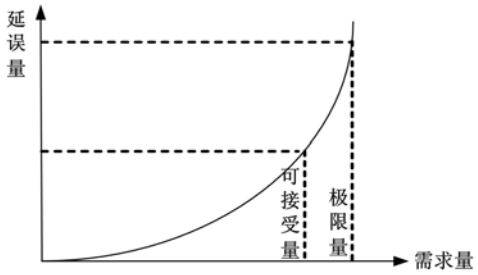


**图 2　航班延误与需求的关系**

**Fig. 2　The relation between flight delay and requirement**

航班延误预警等级一定程度上反映了航班延误所造成的损失，包括本次航班延误所导致的直接损失，以及波及到下游航班所导致的间接损失。但是，想要准确衡量延误造成的损失，需要用大量信息进行复杂计算，一种简单的方法就是用航班延误时间或延误数量来衡量。实际上航班延误数量和延误时间关系密切，一般情况下，延误数量大会导致延误时间长，而延误时间长也会导致延误数量大。本文根据航班延误数量来划分五色预警等级，如表 1 所示，延误程度相对于最大延误架次数而言，预警颜色越深表示机场发生大面积延误的可能性越大。

表 1 机场航班延误五色预警等级

Table 1 Three-Colors warning level of flight delays

| 延误等级 | 五色等级 | 航班延误程度 |
|---|---|---|
| 1 级 | 绿 | 0－10% |
| 2 级 | 蓝 | 11%－20% |
| 3 级 | 黄 | 21%－40% |
| 4 级 | 橙 | 41%－60% |
| 5 级 | 红 | 61%－100% |

从这样的划分不难看出,对不同时段延误的预警实质是个多类分类问题,关键是判定其所属等级。本文将先验知识融合到多类 CVM,并转化为中心约束 MEB 来解决五级延误分类问题,并与原始方法进行了比较。

## 2 基于加权多类 CVM 算法的航班延误预警

机场航班延误预警主要是通过对航班延误及其波及进行识别、分析与评价,向航空公司运行指挥中心、机场运行指挥中心、机场离港系统、旅客等提出警示的过程,具体包括:监测、识别、诊断与评价 4 个阶段。本文将基于空运需求与机场容量冲突导致航班延误的观点,针对大样本学习问题,提出基于加权多类 CVM 算法来对机场航班延误预警分析。

### 2.1 中心约束 MEB 问题

给定一点集 $S=\{x_i\}_{i=1}^{l}$,其中 $x_i\in R^d$,$d$ 是维数,集合 S 的 MEB[8] 是指包含集合 S 中所有数据点的最小球,表示为 $MEB(S)$。如图 3 所示,当 $MEB(S)$ 的中心点为 $c^*$,半径为 $R^*$ 时,最小闭包球 $MEB(S)$ 可以表示为 $B(c^*,R^*)$。令 $k$ 表示映射 $\varphi: x\rightarrow\varphi(x)$ 的核函数。MEB 问题是找一个最小的球包含在特征空间中所有的点 $\varphi(x_i)\in S$。接下来我们讨论中心约束 MEB 问题[9]。对每一个 $\varphi(x_i)$ 首先增加一个额外的项 $\delta_i\in R$,形成 $[\varphi(x_i)\ \delta_i]^{\mathrm{T}}$。同时也约束球心的最后一维为 0(即 $[c\ 0]^{\mathrm{T}}$),此问题的原始形式为

$$\begin{cases}\min\limits_{R,c} R^2\\ \text{s.t.}\quad \|c-\varphi(x_i)\|^2+\delta_i^2\leqslant R^2,\ i=1,\cdots,m\end{cases}\tag{1}$$

相应的对偶形式为

$$\begin{cases}\max\ \boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(K)+\boldsymbol{\Delta})-\boldsymbol{\alpha}^{\mathrm{T}}K\boldsymbol{\alpha}\\ \text{s.t.}\quad \boldsymbol{\alpha}^{\mathrm{T}}1=1,\boldsymbol{\alpha}\geqslant 0\end{cases}\tag{2}$$

式中:$\boldsymbol{\Delta}=[\delta_1^2,\cdots,\delta_m^2]^{\mathrm{T}}\geqslant 0$。

我们能够得到

$$R=\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(K)+\boldsymbol{\Delta})-\boldsymbol{\alpha}^{\mathrm{T}}K\boldsymbol{\alpha}}$$

$$c=\sum_{i=1}^{m}\boldsymbol{\alpha}_i\varphi(x_i)$$

因为式(2)的约束条件 $\boldsymbol{\alpha}^{\mathrm{T}}1=1$,选取一个任意的实数 $\eta\in R$,$\eta$ 倍数的 $\boldsymbol{\alpha}^{\mathrm{T}}1=1$ 能够被增加到目标中时不影响极大值下的参数求解。即式(2)等价于

$$\begin{cases}\max\ \boldsymbol{\alpha}^{\mathrm{T}}(\mathrm{diag}(K)+\boldsymbol{\Delta}-\eta_1)-\boldsymbol{\alpha}^{\mathrm{T}}K\boldsymbol{\alpha}\\ \text{s.t.}\ \boldsymbol{\alpha}^{\mathrm{T}}1=1,\boldsymbol{\alpha}\geqslant 0\end{cases}\tag{3}$$

任何式(3)形式的 QP(二次规划)问题,满足 $\boldsymbol{\Delta}\geqslant 0$ 时能够被转化为中心约束 MEB 问题。注意式(3)不要求核函数一定要满足 $k(x,x)=\kappa$ 即核函数对角线元素为同一常数这一限制条件。而中心约束 MEB 问题可以用文献[12]提出的 $(1+\varepsilon)$ 近似闭包球算法来很好地解决,如图 4 所示。

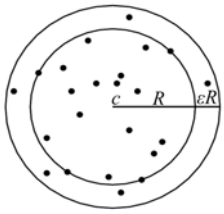


图 3 最小闭包球 MEB

Fig. 3 $MEB(S)=B(c^*,R^*)$

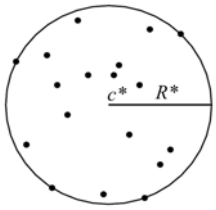


图 4 近似闭包球

Fig. 4 $B(c,(1+\varepsilon)R)$

### 2.2 加权的多类

多类 SVM 可以看成输出是标签向量的 SVM,这个想法来源于用超平面来分离向量的方法[13]。这个向量能够被看作一个特征向量映射到一个一维子空间的映射算子,这些映射排列就可以扩展到多维子空间,这样就解决了标签向量的训练方法。

#### 2.2.1 融合先验知识到多类 SVM

设训练数据集 $s=\{(\boldsymbol{x}_i,\boldsymbol{y}_i,\vec{\boldsymbol{v}}_i)\}_{i=1}^{m}$,其中 $\boldsymbol{x}_i\in R^d$,$\boldsymbol{y}_i,\vec{\boldsymbol{v}}_i\in R^{\mathrm{N}}$,N 为类别数。定义多类标签向量[10] $\boldsymbol{y}_i\in R^{\mathrm{N}}$,记 $\boldsymbol{y}_i=[(y_i)_1,\cdots,(y_i)_{\mathrm{N}}]^{\mathrm{T}}$,其中 $(\boldsymbol{y}_i)_t$ 表示多类标签向量 $\boldsymbol{y}_i$ 的第 $t$ 个分量

$(\mathbf{y}_i)_t = \sqrt{(N-1)/N}$;如果 $x_i$ 属于 $t$ 类

$(\mathbf{y}_i)_t = \sqrt{1/(N(N-1))}$;其他

相应地,对应于每个 $\mathbf{y}_i$ 有置信值向量 $\vec{v}_i = [(\vec{v}_i)_1,\cdots,(\vec{v}_i)_N]^T$,其中 $(\vec{v}_i)_t \in (0,1]$ 表示第 $t$ 类的置信水平。直觉上,样本点属于某个类的置信值越大,我们越希望它属于这一类。

**定义 1**　加权样本点 $(x_i, y_i, \vec{v}_i)$ 关于一个超平面 $(W,\vec{b})$ 的权值函数为

$$(\vec{f}(\vec{v}_i)\odot y_i)^T(W\varphi(x_i)+\vec{b})$$

式中:$\vec{f}:[(\vec{v}_i)_1,\cdots,(\vec{v}_i)_N]^T \to [f((\vec{v}_i)_1),\cdots, f((\vec{v}_i)_N))^T$;$f(\cdot)\in(0,1]$ 是单调减函数,实际中可以令 $f(x)=1/x$;$\odot$ 表示 Hadamard product。

**定义 2**　给定目标边界 $\gamma>0$ 和边界标准函数 $\vec{f}$,定义加权样本点 $(x_i, y_i, \vec{v}_i)$ 关于超平面 $(W,\vec{b})$ 的权值边界松弛变量 $\xi_i^w$ 如下

$$\xi_i^w = \max[0,\gamma-[\vec{f}(\vec{v}_i)\odot y_i]^T[W\varphi(x_i)+\vec{b}] = \xi_i$$

**性质 1**　给定在特征空间线性可分的加权样本点集 $s=\{(x_i, y_i, \vec{v}_i)\}_{i=1}^m$,其中 $x_i\in R^d$;$y_i, \vec{v}_i \in R^N$,下面运用超平面 $(W,\vec{b})$ 解决加权多类优化问题。

加权的多类问题的原始优化形式为

$$\min_{W,b,\rho,\xi_i} \text{trace}(W^TW)+\|\vec{b}\|^2-2\rho+\frac{1}{vm}\sum_{i=1}^m\xi_i^2$$
$$\text{s.t.}\quad (\vec{f}(\vec{v}_i)\odot y_i)^T(W\varphi(x_i)+\vec{b})\geqslant\rho-\xi_i \tag{4}$$

相应的对偶形式为

$$\begin{cases}\min\limits_{\alpha}\sum\limits_{i,j=1}^m\alpha_i\alpha_j(\langle\vec{f}(\vec{v}_i)\odot y_i,\vec{f}(\vec{v}_j)\odot y_j\rangle k(x_i,x_j)+\\ \langle\vec{f}(\vec{v}_i)\odot y_i,\vec{f}(\vec{v}_j)\odot y_j\rangle+\delta_{ij}vm)\end{cases}$$
$$\text{s.t.}\sum_{i=1}^m\alpha_i=1,\alpha_i\geqslant0,\forall\, i \tag{5}$$

从 KKT 条件解出

$$W=\sum_{i=1}^m\alpha_i(\vec{f}(\vec{v}_i)\odot y_i)\varphi(x_i)^T$$
$$\vec{b}=\sum_{i=1}^m\alpha_i(\vec{f}(\vec{v}_i)\odot y_i)$$

根据超平面 $(W,\vec{b})$,下面的决策函数可以预测给定测试样本 $x_i\in R^d$ 的类别

$$\arg\max_{t=1,\cdots,N}\langle y_t,[W\varphi(x_j)+\vec{b}]\rangle =$$
$$\arg\max_{t=1,\cdots,N}\sum_{i=1}^m\{\alpha_i\langle\vec{f}(\vec{v}_i)\odot y_i,y_t\rangle[k(x_i,x_j)+1]\}$$

2.2.2　加权的多类 SVM 转化为 CMEB 问题

为了简化讨论,令 $z_i$ 为 $(x_i,y_i)$,那么样本集 $s=\{z_i\}_{i=1}^m$。下面将加权的多类 SVM 转化为 CMEB 问题,重写公式(5),得

$$\begin{cases}\min\limits_{\alpha}\sum\limits_{i,j=1}^m\alpha_i\alpha_j\tilde{K}(z_i,z_j)\\ \text{s.t.}\sum\limits_{i=1}^m\alpha_i=1,\alpha_i\geqslant0,\forall\, i\end{cases} \tag{6}$$

式中:

$$\tilde{K}(z_i,z_j)=[\langle\vec{f}(\vec{v}_i)\odot y_i,\vec{f}(\vec{v}_j)\odot y_j\rangle k(x_i,x_j)+\langle\vec{f}(\vec{v}_i)\odot y_i,\vec{f}(\vec{v}_j)\odot y_j\rangle+\delta_{ij}vm]$$

令 $\eta=\max_i\{\tilde{K}_{ii}\}$,$\Delta=-\text{diag}(\tilde{K})+\eta_1$,易知 $\Delta\geqslant0$,则式(6)可以写成

$$\max-\boldsymbol{\alpha}^T\tilde{K}\boldsymbol{\alpha} \tag{7}$$
$$\text{s.t.}\ \boldsymbol{\alpha}^T1=1,\boldsymbol{\alpha}\geqslant0$$

即加权的多类 SVM 能够转化为中心约束的 MEB 问题。然后用 Badoiu 和 Clarkson[12] 的技术可以很好地解决中心约束的 MEB 问题。

## 3　试　验

### 3.1　预警模型的建立

以首都机场 2006 年的航班进离港信息表为原始数据,该表提供了按时间顺序记录的进、离港航班的计划起降时间、实际起降时间、跑道及其他相关信息。根据这些信息就能计算出每个航班的延误时间,进而分跑道、分进离累计出单位时段的延误架次。由于要对进港延误和离港延误分别预警,因而要从中提取两个数据集。以离港延误数据集为例,它包含以下特征:跑道 1(2)的计划离港航班架次、延误离港架次、上时段离港延误遗留架次。最后根据表 1 定义的预警等级为每个样本计算延误程度,标记所属类。表 2 给出了离港延误各等级样本占总数的比例。限于篇幅,预警模型的建立过程仅以离港延误预警为例。

**表 2　离港延误各等级样本比例**

**Table 2　Percentage of samples of each grade**

| 等级 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 比例 | 37% | 30% | 17% | 13% | 3% |

### 3.2　样本权值的计算

给定训练数据集为 $s=\{(x_i,y_i)\}_{i=1}^m$,其中 $x_i$

$\in R^d$, $y_i \in R^N$, 对一些整数 $d, N>0$, 假设有 $m$ 个训练样本点分为 N 类, 置信值向量 $\vec{v}_i = [(\vec{v}_i)_1, \cdots, (\vec{v}_i)_N]^T$, 其中 $(\vec{v}_i)_t \in (0,1]$ 表示 $x_i$ 属于 $t$ 类的置信水平。

根据统计分析可知, 样本离某类中心越近则获得该类标记的可能性越大, 常见的基于距离的置信值计算方法为

$$(\vec{v}_i)_t = 1 - \frac{d_t(x_i)}{R_t} + \delta$$
$$y_i = j, t = 1,2,3,4,5 \qquad (8)$$

式中: $d_t(x_i)$ 为 $x_i$ 到 $t$ 类样本中心的距离; $R_t = \max_i d(x_i)$ 为 $t$ 类样本的类半径; $\delta>0$ 为一个很小的常数, 避免出现权值为 0 的情况。

该方法根据样本到 $j$ 类中心的距离衡量样本为 $j$ 类的权值, 只考虑了类内的样本紧密度, 所计算的权值是绝对的, 位于类边缘的样本只能获得很低的权值。在二分类情况下, 有些边缘样本属于正类的可能性远大于属于负类的可能性, 如图 5 所示, 样本 a 和样本 b 均位于正类边缘, 有着相同的权值, 且取值很小, 但 a 到负类中心的距离远大于到正类中心的距离, 因而, 相对于负类而言, a 属于正类的可能性更大; 相对于 b 而言, a 属于正类的权值应当更大。

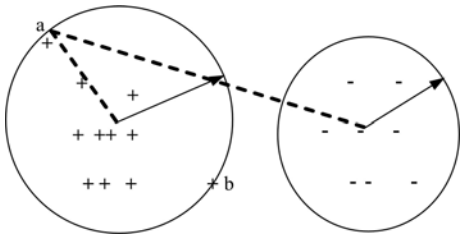


**图 5 绝对紧密度与相对紧密度**

**Fig. 5 The relative membership of samples**

针对上述情况, 本文提出了一种基于相对距离的权值计算方法, 同时考虑样本与各类中心的距离, 用相对的紧密度来衡量所属分类的可能性。结合提出的五级预警模型, 这些样本的置信值向量 $\vec{v}_i = [(\vec{v}_i)_1, \cdots, (\vec{v}_i)_N]^T$ (其中 $N=5$)可用下式计算

$$(\vec{v}_i)_t = \sum_{j=1}^{N} \{0.5 + \frac{\exp[C_0[d_j(x_i) - d_t(x_i)]/d_{j,t}] - \exp(-C_0)}{2[\exp C_0 - \exp(-C_0)]}\} \qquad (9)$$

式中: $d_j(x_i)$ 为 $x_i$ 到 $j$ 类样本中心的距离; $d_t(x_i)$ 为 $x_i$ 到 $t$ 类样本中心的距离; $d_{j,t}$ 为 $j$ 类样本中心到 $t$ 类样本中心的距离; $C_0$ 为控制参数, 计算时需要提前给定。

再对由上式求出的置信值向量 $\vec{v}_i = [(\vec{v}_i)_1, \cdots, (\vec{v}_i)_N]^T$ 进行标准化: $\vec{v}_i = \left[\frac{(\vec{v}_i)_1}{\|\vec{v}_i\|}, \cdots, \frac{(\vec{v}_i)_N}{\|\vec{v}_i\|}\right]^T$。这样既能保证靠近类中心的样本有大的权值, 也能保证多数边缘样本有较大权值, 而取得小权值的样本正是处于两类交界处的样本。显然, 用该方法算得的权值更符合样本的分布情况。

### 3.3 控制参数的筛选

表 3 是加权算法中控制参数 $C_0$ 对 WMMCVM 最终分类性能的影响, WMMCVM 代表我们提出的加权多类 CVM, 样本集为某机场一年的数据, 大小为 31 870 条。

**表 3 控制参数 $C_0$ 对分类性能的影响**

**Table 3 Influence of control value**

| $C_0$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 总分类正确率/% | 85.3 | 87.5 | 90.1 | 82.2 |

各级分类器均采用径向基核函数(RBF), C 取 100, $\varepsilon = 10^{-3}$, 用 5 次交叉验证法训练和验证, 总分类正确率是指所有类的总正确率。结果表明 $C_0 = 3$ 时, 最终分类效果最好。后面试验都选 $C_0 = 3$ 来确定权值

表 4 是控制参数 $\varepsilon$ 对最终分类性能的影响, 参数选择如上例, 结果表明 $\varepsilon \geqslant 10^{-3}$ 时, 最终分类效果较好, 并且训练时间又比较合适。为了得出好的分类性能且快的训练速度, 后面实验都选取 $\varepsilon = 10^{-3}$。

**表 4 控制参数 $\varepsilon$ 对分类性能的影响**

**Table 4 Influence of control value $\varepsilon$**

| $\varepsilon$ | $10^{-1}$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ | $10^{-7}$ |
|---|---|---|---|---|---|---|---|
| 时间/s | 54 | 117 | 238 | 483 | 936 | 1817 | 3109 |
| 正确率/% | 78.4 | 85.8 | 88.1 | 89.6 | 90.1 | 91.7 | 92.9 |

### 3.4 比较

为了进行模型性能的对比, 分别训练了基于 MCVM[10], OAO-WMSVM 及本文的 WMMCVM 的五级预警模型。WMMCVM 代表作者提出的加权多类 CVM, OAO-WMSVM 代表一对一加权间隔支持向量机, MCVM 代表多类 CVM。样本集为某机场一年的数据, 大小为 31 870 条。各级分类器均采用径向基核函数(RBF), C取100, 用5次交叉验证法训练和验证。单调递减函数 $f(x) = 1/x$。表 5 给出了各预警模型对各等级延误的预报

准确率，表 6 为各级分类模型的训练时间对比图。

试验结果表明，融合权值的 OAO-WMSVM 和本文的 WMMCVM 的各等级延误预报能力明显优于未加权的 MCVM 预警模型，说明融合先验知识确实能有效提高 SVM 的分类能力。虽然 WMMCVM 的各等级延误预报能力略优于一对一加权间隔支持向量机（OAO-WMSVM），但训练速度却明显优于后者。

表 5 中 3 级延误预报准确率最低。分析样本可以找到原因：3 级样本的空间分布较为杂乱，分类难度大，而 4、5 级样本却有较紧凑的分布，因而更容易区分。基于 WMMCVM 的航班延误预警模型的重要价值在于：对发生最频繁的 1、2 级延误及较严重的 4、5 级延误都有较高的预报准确率。

**表 5　各级分类模型的分类能力(%)**

**Table 5　Performance of classifiers(%)**

| 等级 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| MCVM | 88.1 | 77.2 | 69.1 | 82.4 | 82.4 |
| OAO-WMSVM | 91.7 | 83.1 | 76.1 | 92.1 | 92.1 |
| WMMCVM | 93.3 | 84.4 | 79.6 | 94.9 | 95.6 |

**表 6　各级分类模型的训练时间(s)**

**Table 6　Training time of classifiers in seconds**

| | |
|---|---|
| MCVM | 127 |
| OAO－WMSVM | 3172 |
| WMMCVM | 241 |

## 4　结束语

针对进离港航班需求量与机场容量之间的矛盾导致航班延误问题，用 WMMCVM 方法建立了一个五级航班延误预警模型。该方法能利用样本权值将先验知识融合到分类器中，通过权值函数调节不同样本对决策分类面的贡献，得到最优分类面。试验表明，用这种融合了样本分布知识的航班延误预警模型具有很好的预报能力和非常快的训练速度。

**参考文献：**


[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. 2nd Ed. New York:Springer, 2000.

[2] Wu X, Srihari R. Incorporating prior knowledge with weighted margin support vector machines[C]// In the 10th ACM SIGKDD International Conference on Knowledge, 2004:326-333.

[3] Fabien L, Gerard B. Incorporating prior knowledge in support vector machine for classification: A review [J]. Neurocomputing, 2008, 71(7-9):1578-1594 .

[4] Schölkopf, Simard P, Smola A, et al. Prior knowledge in support vector kernels[C]// In Kernel Methods-support Vector Learning, MA, USA: MIT Press, 1998:640-646.

[5] Fung G, Mangasarian O L, Shavlik J. Knowledge-based support vector machine classifiers[R]. In Data Mining Institute Technical Report 01-09, 2001.

[6] Platt J C. Fast training of support vector machines using sequential minimal optimization[C]// In kernel Methods-support Vector Learning, Cambridge, MA, USA: MIT Press, 1999:185-208.

[7] Collobert R, Bengio S, Bengio Y. A parallel mixture of SVMs for very large scale problems[J]. Neural Computation, 2002, 14: 1105-1114.

[8] Tsang I W, Kwok J T, Cheung P M. Core vector machines: Fast SVM training on very large data sets [J]. Journal of Machine Learning Research, 2005, 6: 363-392.

[9] Tsang I W, Kwok J T, Lai K T. Core vector regression for very large regression problems[C]// In the Twenty-Second International Conference on Machine Learning. 2005:913-920.

[10] Asharaf S, Murty M N, Shevade S K. Multiclass core vector machine[C]// In the 24th International Conference on Machine Learning, ACM, 2007:41 - 48.

[11] Malone K M. Modeling a network of queues under nonstationary and stochastic conditions[D]. Cambridge: Sloan School of Management, Massachusetts Institute of Technology, 1993.

[12] Baidoiu M, Clarkson K L. Optimal core sets for balls [C]// In DIMACS Workshop on Computational Geometry, 2002.

[13] Szedmak S, Taylor J S. Multiclass learning at one-class complexity[R]. Technical Report No: 1508, School of Electronics and Computer Science, Southampton, UK, 2005.