一族三次样条光滑半监督支持向量机

王建建 张晓丹
（北京科技大学数理学院 北京 100083）

摘 要 针对半监督支持向量分类优化中的非凸非光滑化问题，建立光滑半监督支持向量机模型，提出基于分段多项式函数和插值思想构造一个新的三次样条光滑函数，从而可以更好地逼近对半监督支持向量机中非光滑的对称铰链损失函数部分，构造出基于此光滑函数的具有二阶光滑的半监督支持向量机模型。进而可以用优化中的光滑算法来求解该模型，并分析所构造的三次样条函数对对称铰链损失函数的逼近精度。通过数据实验证明所构造的新的光滑半监督模型具有较好的分类效果和效率。

关键词 半监督支持向量机 三次样条函数 铰链光滑 分类

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2015.08.011

A CUBIC SPLINE SMOOTH SEMI-SUPERVISED SUPPORT VECTOR MACHINE

Wang Jianjian Zhang Xiaodan
(*School of Mathematic and Physics, Beijing University of Science and Technology, Beijing 100083, China*)

Abstract For the problem of non-convex non-smooth in semi-supervised support vector classification optimisation, we built a smooth semi-supervised support vector machine (SVM) model, and proposed to construct a new cubic spline smooth function based on piecewise polynomial function and interpolation idea, so as to better approach the non-smooth symmetric hinge loss function part in semi-supervised SVM, and construct the semi-supervised SVM model which is based on this smooth function and has two-order smooth. Furthermore the smooth algorithm in optimisation can be used to solve the model. We also analyse the approaching accuracy of the constructed cubic spline function on symmetric hinge loss function. Through data experiment it is proved that the built new smooth semi-supervised SVM model has better classification effect and efficiency.

Keywords Semi-supervised SVM Cubic spline function Hinge smooth Classification

0 引 言

半监督学习中，已标识的样本数较少在现实中是普遍存在的，仅利用这些样本提供的信息不能有效地训练出好的符合要求的分类界面。因此人为地加入未标识样本，采用某种准则将其假设为标识样本来提高算法的分类能力。然而在算法优化阶段又存在非凸非光滑的难题，半监督中存在的这些问题的研究具有巨大利用价值，尽管半监督支持向量机算法的研究进度迅猛，但是在样本信息的获取和预处理阶段还是存在巨大的问题，因此半监督应用前景引起了越来越多研究人员的关注[1-3]。

半监督支持向量机模型把间隔最大化的原则应用到标号和未标号的样本中，因而不像支持向量机最后得到的是一个凸优化问题，半监督支持向量机得到的是一个非凸非光滑的问题。2005 年 Chapelle 等建立一个无约束的光滑半监督支持向量机模型，并用光滑的高斯近似函数来逼近无标记样本的对称铰链损失函数[4]，使原来非光滑模型变成光滑模型，进而用优化算法来求解。2009 年刘叶青等人建立了多项式光滑的半监督支持向量机模型，引入一族多项式光滑函数来逼近非凸的目标函数[5]。光滑函数使得原本不可微的半监督模型变成可微的模型，从而可以采用快速优化的算法来求解，极大程度的降低了半监督支持向量机的计算复杂度。本文应用分段多项式函数和插值思想构造了三次样条光滑函数来逼近对称铰链损失函数，建立了光滑的半监督支持向量机模型，同时研究并分析了此光滑函数的性质及其逼近精度分析并进行数值实验。

1 光滑半监督支持向量机模型

考虑支持向量机的二分类问题[6,7,11]，训练集包括 m 个标号样本 $\{(x_i,y_i)\}_{i=1}^{m}$ 和 l 个未标号样本 $\{x_i\}_{i=m+1}^{m+l}$。其中，$x_i \in R^n$ 为行向量，$y_i \in \{1,-1\}$。将上述 m 个标记样本 $x_i(i=1,2,\cdots,m)$ 用矩阵 $A_{m\times n}$ 表示。$x_1,x_2,\cdots,x_m$ 被分为 A^+ 和 A^- 两类，若 x_i 属于类 A^+，记为 1，若 x_i 属于类 A^-，记为 -1；这样，可以用一个 $m\times m$ 的对角阵 D 来表示分类情况，D 的对角元素为 1 或 -1。而对 l 个未标记样本点，用矩阵 $B_{l\times n}$ 表示。记 $e_i(i=1,2)$ 为分量全为 1 的列向量，$e_1 \in R^m, e_2 \in R^l$，$\omega \in R^n, b \in R$。研究如下半监督支持向量机模型：

$$\begin{aligned} &\min_{\omega,b} \frac{1}{2}\|\omega\|_2^2 + ce_1^T\xi_1 + c^*e_2^T\xi_2 \\ &\text{s.t} \quad D(A\omega + e_1b) \geqslant e_1 - \xi_1 \quad \xi_1 \geqslant 0 \\ &\quad |B\omega + e_2b| \geqslant e_2 - \xi_2 \quad \xi_2 \geqslant 0 \end{aligned} \tag{1}$$

收稿日期：2014-04-14。国家自然科学基金项目(70271068)。王建建，硕士生，主研领域：支持向量机。张晓丹，教授。

这里，c 和 c^* 是错分惩罚参数，ξ_1,ξ_2 为松弛向量，$\xi_1 \in R^m,\xi_2 \in R^l$。令 ξ_i 为如下形式：

$$\xi_1 = L(D(A\omega + e_1 b)),\xi_2 = L(|B\omega + e_2 b|) \quad (2)$$

其中，$L(x) = \max(0,1-x)$ 为铰链损失函数，$L(|x|) = \max(0,1-|x|)$ [8] 为对称铰链损失函数。若 $x \in R^n$，则令 $L(x) = (L(x_1),L(x_2),\cdots L(x_n))^T$ 将式(1)目标函数中的 $\xi_i(i=1,2)$ 用式(2)代替，得到无约束优化问题：

$$\min_{\omega,b} \frac{1}{2}\|\omega\|_2^2 + ce_1^T L(D(A\omega + e_1 b)) + c^* e_2^T L(|B\omega + e_2 b|) \quad (3)$$

由于铰链损失函数与对称铰链损失函数均是不可微的，它使得无约束优化模型式(3)为非光滑规划问题。

本文对模型式(3)进行修正得到如下半监督支持向量机模型：

$$\min_{\omega,b} \frac{1}{2}\|\omega\|_2^2 + \frac{c}{2}\|L(D(A\omega + e_1 b))\|_2^2 + \frac{c^*}{2}\|L(|B\omega + e_2 b|)\|_2^2 \quad (4)$$

这种修正对原问题的影响很小，但却能避开铰链损失函数的不可微性以及避开对称铰链损失函数的两个不可微点。基于式(4)，建立如下光滑半监督支持向量机：

$$\underset{(\omega^T,b)\in R^{n+1}}{\varphi(\omega,b)} = \min_{(\omega^T,b)\in R^{n+1}} \frac{1}{2}\|\omega\|_2^2 + \frac{c}{2}\|L(D(A\omega + e_1 b))\|_2^2 + \frac{c^*}{2}\|f(|B\omega + e_2 b|)\|_2^2 \quad (5)$$

其中，$f(x)$ 为逼近对称铰链损失函数的任一光滑函数。本文研究一族三次样条光滑函数逼近对称铰链损失函数的光滑半监督支持向量机。

2 样条函数模型

2.1 构造三次样条函数

样条函数是函数逼近中一个十分活跃的一部分[9-10]，适用性比较广，而且类型也比较多。本节推导如下形式的三次样条函数：

$$S_3(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \sum_{i=1}^{n-1} C_i (x - x_j)_+^3 \quad (6)$$

逼近对称铰链损失函数。式中，操作符 $(\cdot)_+^m$ 的定义如下：

$$(u)_+^m = \begin{cases} u^m & u > 0 \\ 0 & u \leqslant 0 \end{cases} \quad (7)$$

定理 1 设 $k > 1, x_0 = -1/k, x_1 = 0, x_2 = 1/k$ 是节点，则存在唯一一个逼近 $L(|x|)$ 的三次样条插值函数 $S_3(x,k)$ 满足条件：

$$S_3\left(\pm\frac{1}{k},k\right) = 1 - \frac{1}{k} \quad S_3(0,k) = 1$$

$$S'_3(x_0) = 1 \quad S'_3(x_2) = -1$$

证明：

三个节点，所以 $n = 2$。此时，式(6)为如下形式：

$$\begin{aligned} S_3(x,k) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \sum_{i=1}^{2-1} C_i (x - x_j)_+^3 \\ &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + C_1 (x - x_1)_+^3 \\ &= \begin{cases} a_0 + a_1 x + a_2 x^2 + a_3 x^3 & x \leqslant x_1 \\ a_0 + a_1 x + a_2 x^2 + a_3 x^3 + C_1 (x - x_1)_+^3 & x > x_1 \end{cases} \end{aligned} \quad (8)$$

通过定理的条件计算可以得到：$a_0 = 1, a_1 = 0, a_2 = -2k, a_3 = -k^2, C_1 = 2k^2$，所以逼近对称铰链损失函数 $L(|x|)$ 的二阶光滑的三次样条函数为：

$$S_3(x,k) = \begin{cases} 1 + x & x \leqslant -\frac{1}{k} \\ 1 - 2kx^2 - k^2 x^3 & -\frac{1}{k} < x \leqslant 0 \\ 1 - 2kx^2 + k^2 x^3 & 0 < x < \frac{1}{k} \\ 1 - x & x \geqslant \frac{1}{k} \end{cases} \quad (9)$$

该函数逼近对称铰链损失函数的效果如图(1)所示。

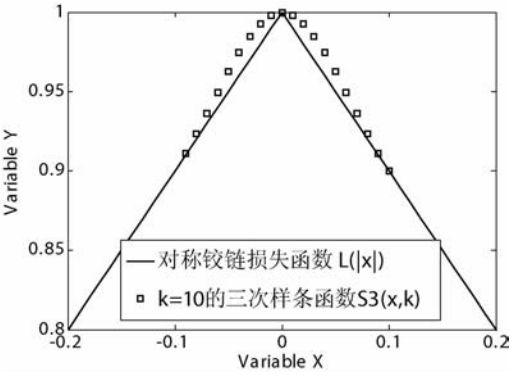

图 1 逼近对称铰链损失函数 $L(|x|)$ 的三次样条函数 $S_3(x,k)$

定理 2 $S_3(x,k)$ 为式(9)定义的三次样条函数，$L(|x|)$ 为对称铰链损失函数，则有：

(1) $S_3(x,k)$ 关于 x 二阶光滑；

(2) $S_3(x,k) \geqslant L(|x|)$；

(3) $S_3(x,k) - L(|x|) \leqslant \frac{4}{27k}$，且在 $x = \pm\frac{1}{3k}$ 相等。

证明：

(1) 可直接验证在 $x = 0$ 和 $x = \pm\frac{1}{k}$ 处满足以下条件：

$$S_3\left(-\frac{1}{k},k\right) = 1 - \frac{1}{k} \quad S'_3\left(-\frac{1}{k},k\right) = 1$$

$$S_3\left(\frac{1}{k},k\right) = 1 - \frac{1}{k} \quad S'_3\left(\frac{1}{k},k\right) = -1$$

故结论(1)成立。

(2) 在 $x \leqslant -\frac{1}{k}$ 和 $x \geqslant \frac{1}{k}$ 上，$S_3(x,k) - L(|x|) = 0$，结论成立。样条函数在区间 $\left(-\frac{1}{k},0\right]$ 上，$S_3(x,k) - L(|x|) = -x(kx+1)^2 \geqslant 0$；在区间 $\left(0,\frac{1}{k}\right)$ 上，$S_3(x,k) - L(|x|) = x(kx+1)^2 \geqslant 0$，式中 $k > 1$ 故结论成立。

(3) 样条函数式(9)在区间 $\left(-\frac{1}{k},0\right]$ 满足条件 $S_3(x,k) \geqslant L(|x|)$，且先增大后减小，并且在 $x = -\frac{1}{3k}$ 处取得最大，$S_3(x,k) - L(|x|) \leqslant S_3\left(-\frac{1}{3k},k\right) = \frac{4}{27k}$。

在区间 $\left(0,\frac{1}{k}\right)$ 上有 $S_3(x,k) \geqslant L(|x|)$，令：

$$y(x) = S_3(x,k) - L(|x|) = x - 2kx^2 + k^2 x^3$$

对 $y(x)$ 求导可得 $y'(x) = 1 - 4kx + 3k^2 x^3 = (kx-1)(3kx-1)$ 最大值在 $x = \frac{1}{3k}$ 处取得，$y\left(\frac{1}{3k}\right) = S_3\left(\frac{1}{3k},k\right) - L\left(\left|\frac{1}{3k}\right|\right) = x - 2kx^2 + k^2 x^3 = \frac{4}{27k}$。

故结论 $S_3(x,k)-L(|x|)\leqslant\frac{4}{27k}$,且在 $x=\pm\frac{1}{3k}$ 相等成立。

2.2　一族三次样条函数

根据定理 2,$S_3(x,k)$ 整体向下移动 $\frac{4}{27k}$,在区间 $\left(-\infty,\frac{1}{3k}\right]$ 和 $\left[\frac{1}{3k},\infty\right)$ 上依然取对称铰链损失函数 $L(|x|)$,可得到逼近 $L(|x|)$ 的一个新的光滑的三次样条函数:

$$f_3(x,k)=\begin{cases}1+x & x\leqslant-\frac{1}{3k}\\ 1-2kx^2-k^2x^3-\frac{4}{27k} & -\frac{1}{3k}<x\leqslant 0\\ 1-2kx^2+k^2x^3-\frac{4}{27k} & 0<x<\frac{1}{3k}\\ 1-x & x\geqslant\frac{1}{3k}\end{cases}\tag{10}$$

式中 $k>1$。其逼近对称铰链损失的函数图像如图 2 所示。

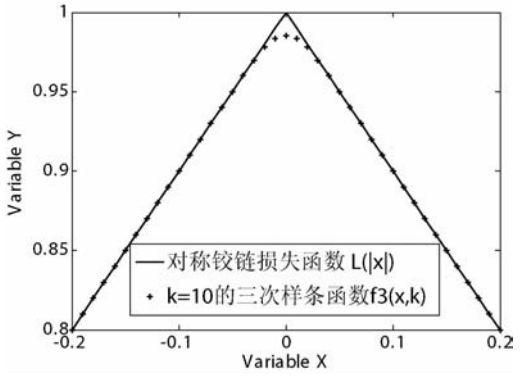

图 2　逼近对称铰链损失函数 $L(|x|)$ 的三次样条函数 $f_3(x,k)$

文献[4]提出用高斯函数 $\tau(x)=e^{-3x^2}$ 来逼近对称铰链损失函数 $L(|x|)$。本文所构造的三次样条函数 $f_3(x,k)$ 与文献中的高斯函数逼近对称铰链函数的效果如图 3 所示。

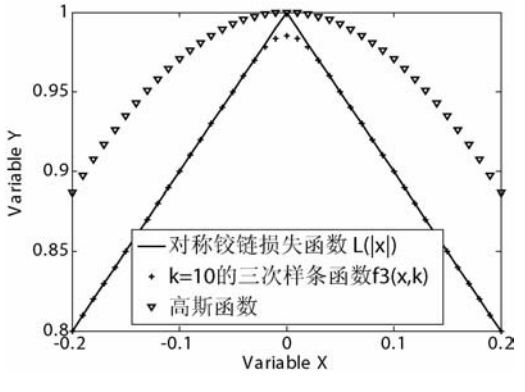

图 3　逼近对称铰链函数的两种不同光滑函数

定理 3　设 $x\in R,k>1$,$L(|x|)$ 是对称铰链损失函数,$f_3(x,k)$ 是式(10)定义的三次样条函数。则有:

(1) $0\leqslant f_3(x,k)\leqslant L(|x|)$;

(2) $0\leqslant L^2(|x|)-f_3^{\,2}(x,k)\leqslant\frac{4}{27k}\left(2-\frac{4}{27k}\right)$。

证明:(1)当 $x\geqslant\frac{1}{3k}$ 或 $x\leqslant-\frac{1}{3k}$ 时,$L(|x|)-f_3(x,k)=0$,结论显然成立。当 $x\in\left(-\frac{1}{3k},0\right)$ 时 $f'_3(x,k)=-4kx-3k^2x^2=-kx(3kx+4)\geqslant 0$,$f_3(x,k)$ 单调递增,故 $f_3(x,k)\geqslant f_3\left(-\frac{1}{3k},k\right)=1-\frac{9}{27k}\geqslant 0$。其次令 $u(x,k)=L(|x|)-f_3(x,k)$,则 $u(x,k)=k^2x^3+2kx^2+x+\frac{4}{27k}$,$u'(x,k)=(kx+1)(3kx+1)\geqslant 0$ 故 $u(x,k)$ 单调递增,在 $x=-\frac{1}{3k}$ 处取得最小值:$u\left(-\frac{1}{3k},k\right)=0$,所以有 $u(x,k)\geqslant 0$,即 $0\leqslant f_3(x,k)\leqslant L(|x|)$。

同理可证当 $x\in\left(0,\frac{1}{3k}\right)$ 时,$0\leqslant f_3(x,k)\leqslant L(|x|)$。

(2)当 $x\geqslant\frac{1}{3k}$ 或 $x\leqslant-\frac{1}{3k}$ 时,结论显然成立。当 $x\in\left(-\frac{1}{3k},0\right)$ 时,由(1)知,$u(x,k)$ 单调递增,在 $x=0$ 处取得最大值:$u(0,k)=\frac{4}{27k}$,$0\leqslant L(|x|)-f_3(x,k)\leqslant\frac{4}{27k}$。又令 $u_2(x,k)=L(|x|)+f_3(x,k)$,则有 $u_2(x,k)=2+x-2kx^2-k^2x^3-\frac{4}{27k}$,$u_2'(x,k)=2-(kx+1)(3kx+1)\geqslant 0$,故 $u_2(x,k)$ 在 $x\in\left(-\frac{1}{3k},0\right)$ 上单调递增,在 $x=0$ 处取得最大值:$u_2(0,k)=2-\frac{4}{27k}$,即 $0\leqslant(L(|x|)+f_3(x,k))\leqslant 2-\frac{4}{27k}$。所以有:$0\leqslant L^2(|x|)-f_3^{\,2}(x,k)=(L(|x|)-f_3(x,k))(L(|x|)+f_3(x,k))\leqslant\frac{4}{27k}\left(2-\frac{4}{27k}\right)$从而结论成立。

同理可证在区间 $x\in(0,\frac{1}{3k})$ 上,亦有:

$0\leqslant L^2(|x|)-f_3^{\,2}(x,k)\leqslant\frac{4}{27k}\left(2-\frac{4}{27k}\right)$成立。

3　三次样条光滑半监督支持向量机模型收敛性分析

定理 4[11]　设 $A\in R^{m\times n}$,$B\in R^{l\times n}$,$x\in R^n$,$\eta\in R^m$,$\mu\in R^l$,$c,c^*\in R$,$k>1$ 定义的实函数如下:

$$h(x)=\frac{1}{2}\|x\|_2^2+\frac{c}{2}\|L(D(Ax+\eta))\|_2^2+\frac{c^*}{2}\|L(|Bx+\mu|)\|_2^2\tag{11}$$

$$g(x,k)=\frac{1}{2}\|x\|_2^2+\frac{c}{2}\|L(D(Ax+\eta))\|_2^2+\frac{c^*}{2}\|f_3(Bx+\mu,k)\|_2^2\tag{12}$$

其中,$f_3(x,k)$ 由式(10)定义。则:

(1) 优化问题 $\min\limits_{x\in R^n}h(x)$ 存在最优解 $\bar{x}$,优化问题 $\min\limits_{x\in R^n}g(x,k)$ 存在最优解 $\bar{x}^k$,并且 $\lim\limits_{k\to\infty}h(\bar{x}^k)=h(\bar{x})$.

(2) 设优化问题 $\min\limits_{x\in R^n}h(x)$ 最优解集合为 D_h,则 $\{\bar{x}^k\}$ 存在收敛子列 $\{\bar{x}^{k_n}\}$ 满足 $\lim\limits_{n\to\infty}\bar{x}^{k_n}=\bar{x}_h$,这里 $\bar{x}_h\in D_h$。

证明:(1)证明方法与文献[11]所提出的方法类似。定义对应的水平集为 $L_\nu(h(x))=\{x|x\in R^n,h(x)\leqslant\nu\}$. $L_\nu(g(x,k))=\{x|x\in R^n,g(x,k)\leqslant\nu\}$。由于 $0\leqslant f_3(x,k)\leqslant L(|x|)$,因此对任意 $\nu\geqslant 0$,它们满足 $L_\nu(h(x))\subset L_\nu(g(x,k))\subset\{x|\|x\|_2^2\leqslant 2\nu\}$。因此 $L_\nu(h(x))$ 和 $L_\nu(g(x,k))$ 是 R^n 空间中的紧集,因此优化问题 $\min h(x)$,$\min g(x,k)$ 的最优解存在,也满足 $\min\limits_{x\in R^n}h(x)=h(\bar{x})$,$\min\limits_{x\in R^n}g(x,k)=g(\bar{x}^k,k)$。其次,对任意 $x\in R^n$,由定理 3 可知,$0\leqslant h(x)-g(x,k)=\frac{c^*}{2}\|L(|Bx+\mu|)\|_2^2-\frac{c^*}{2}\|f_3(Bx+$

$\mu,k)\|_2^2=\frac{c^*}{2}\|L(|Bx+\mu|)\|_2^2-\frac{c^*}{2}\|f_3(Bx+\mu,k)\|_2^2=\frac{c^*}{2}\sum_{i=1}^{i=l}[L^2(|B_ix+\mu_i|)-f_3^2(B_ix+\mu_i,k)]\leqslant\frac{2c^*l}{27k}\left(2-\frac{4}{27k}\right)$所以 $0\leqslant h(\bar{x}^k)-g(\bar{x}^k,k)\leqslant\frac{2c^*l}{27k}\left(2-\frac{4}{27k}\right),0\leqslant h(\bar{x})-g(\bar{x},k)\leqslant\frac{2c^*l}{27k}\left(2-\frac{4}{27k}\right),0\leqslant h(\bar{x})-g(\bar{x},k)\leqslant\frac{2c^*l}{27k}\left(2-\frac{4}{27k}\right)$，又因为 $h(\bar{x}^k)\geqslant h(\bar{x}),g(\bar{x},k)\geqslant g(\bar{x}^k,k)$，所以 $0\leqslant h(\bar{x}^k)-h(\bar{x})\leqslant h(\bar{x}^k)-h(\bar{x})+g(\bar{x},k)-g(\bar{x}^k,k)=h(\bar{x}^k)-g(\bar{x}^k,k)+g(\bar{x},k)-h(\bar{x})\leqslant h(\bar{x}^k)-g(\bar{x}^k,k)\leqslant\frac{2c^*l}{27k}\left(2-\frac{4}{27k}\right)$，从而，$\lim_{k\to\infty}h(\bar{x}^k)=h(\bar{x})$。

（2）对任意 $k\in Z_+,k>1$，由式(10)的定义及定理3的结论(2)可得：$\frac{1}{2}\|\bar{x}^k\|_2^2\leqslant g(\bar{x}^k,k)\leqslant g(\bar{x},k)$，因此，$\{\bar{x}^k\}$ 有界，从而 $\{\bar{x}^k\}$ 有收敛子列 $\bar{x}^{k_n}$。不妨设 $\lim_{n\to\infty}\bar{x}^{k_n}=\bar{x}_h$，可得 $\lim_{n\to\infty}h(\bar{x}^{k_n})=h(\bar{x}_h)=\lim_{k_n\to\infty}h(\bar{x}^{k_n})=h(\bar{x})$，因此，$\bar{x}_h\in D_h$，即 $\bar{x}_h$ 是优化问题 $\min_{x\in R^n}h(x)$ 的最优解。

4 数值试验

实验1 心脏病数据取自 Cleverland Clinic Foundation 基金会，数据样本从 UCI 机器学习数据库可得. 数据样本总量有270个，均为已标记的，将数据随机排序后，取前70个数据为标记数据，对后200个数据做无标记处理. 每个数据样本包含13个属性，所有数据样本被分成两类 presence（有病）和 absence（无病）。

为比较不同光滑半监督支持向量机的分类效果，采用分类器的推广能力作为指标，分类器的推广能力用未标记训练样本的正确率来衡量。实验采用模型式(5)，其中函数 $f(x)$ 分别采取高斯函数和三次样条函数 $f_3(x,k)$。

采用 BFGS-Armijo[12] 算法求解式(12)，求解得到的逼近精度如表1所示。

表1 当 $k=10$ 用 $f_3(x,k)$ 和 e^{-3x^2} 逼近对称铰链损失函数的训练正确率

逼近函数 \ 性能	正确率 $c=c^*=1$	正确率 $c=c^*=3$	正确率 $c=c^*=5$
本文的三次样条函数	84.5	83	82.5
高斯函数	82.5	81.5	81.5

实验2 鸢尾花植物数据集是一个用来检验分类算法性能的标准数据集。此数据集来自 UCI 数据库，该数据集共150个样本点，分为三类：I：Iris-setosa，II：Iris-versicolor 和 III：Iris-virginica，每类样本集各50个样本点，每个样本有4个特征属性，分别为：萼片长度、萼片宽度、花瓣长度、和花瓣宽度，由于是对二类分类进行半监督测试，选取前两类做实验。故取前100个样本点作为训练集，随机打乱顺序，再对前30个数据作为标记数据，后70个数据作为未标记处理，用以上的方法进行训练。如表2所示。

表2 当 $k=10$ 用 $f_3(x,k)$ 和 e^{-3x^2} 逼近对称铰链损失函数的训练正确率

逼近函数 \ 性能	正确率 $c=c^*=5$	正确率 $c=c^*=15$	正确率 $c=c^*=20$
本文的三次样条函数	92.8571	94.2857	94.2857
高斯函数	91.4286	88.5714	84.2857

实验3 Wisconsin Diagnostic Breast Cancer（WDBC）数据，数据样本从 UCI 机器学习数据库可得. 数据有569个样本，共两类均为已标记的，每个样本包含30个属性。将数据随机排序后，取前57个数据为标记数据，对后512个数据做无标记处理.，所有数据样本被分成（M = malignant）和（B = benign）两类。实验采用模型式(5)，其中函数 $f(x)$ 分别采取高斯函数和三次样条函数 $f_3(x,k)$。采用 BFGS-Armijo[12] 算法求解式(12)，求解得到的逼近精度如表3所示。

表3 当 $k=10$ 用 $f_3(x,k)$ 和 e^{-3x^2} 逼近对称铰链损失函数的训练正确率

逼近函数 \ 性能	正确率 $c=c^*=3$	正确率 $c=c^*=10$	正确率 $c=c^*=15$
本文的三次样条函数	86.3281	87.9102	84.1797
高斯函数	85.5234	87.1094	81.0547

实验结果表明，由表1－表3可以得出 $k=10$ 时当 $c=c^*$ 取不同值时，采用本文三次样条函数比高斯函数逼近对称铰链损失函数具有较高的分类精度。因此本文构造的三次样条光滑半监督支持向量机在进行二类分类实验时，利用未标记样本训练时可取的更高的准确率。

5 结 语

本文基于分段多项式函数和插值思想，针对半监督分类模型中非光滑的对称铰链损失函数提出一个定理进行证明得到一个三次样条光滑函数。经过对光滑函数的性质分析得出另一个新的三次样条光滑函数；并对这个新的三次样条光滑函数的逼近精度进行分析并证明。最后进行数值实验并验证了基于本文光滑函数的半监督支持向量分类机可以取得更优越的分类精度。

参 考 文 献

[1] 邓乃扬，田英杰. 数据挖掘的新方法—支持向量机[M]. 北京：科学出版社，2004：288－355.

[2] Chapelle O，Sindhwani V，Keerhi S S，et al. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research，2008，9(2)：203－233.

[3] Reddy S，Shevade S，Murty M N. A fast quasi-Newton method for semi-supervised svm[J]. Pattern Recognition，2011：2305－2313.

[4] Chapelle O，Zien A. Semi-supervised classification by low density separation [C]//Proceedings of 10th international Workshop on AI&Statistics，2005：179－181.

[5] 刘叶青，刘三阳，谷明涛. 一种多项式光滑的半监督支持向量机分类算法[J]. 计算机科学，2009，36(7)：179－181.

[6] Lee Y J，Mangasarian O L. SSVM：A smooth support vector machine for classification[J]. Computational Optimization and Applications，2001，22(1)：5－21.

[7] 袁玉波,严杰,徐成贤. 多项式光滑的支撑向量机[J]. 计算机学报,2005,28(1):9-17.
[8] 吴青. 基于最优化理论的支持向量机学习算法研究[D]. 西安电子科技大学,2009.
[9] Ahlberg J H, Nilson E N, Walsh J L. The theory of splines and their application[M]. New York: Academic Press, 1967.
[10] 袁华强,涂文根,熊金志. 一个新的多项式光滑支持向量机[J]. 计算机科学,2011,38(3):243-247.
[11] 马菁改. 基于广义样条函数的光滑半监督支持向量机的研究与应用[D]. 北京:北京科技大学数理学院,2013.
[12] 熊金志,袁华强,彭宏. 多项式光滑的支持向量机一般模型研究[J]. 计算机研究与发展,2008,45(8):1346-1353.

(上接第 16 页)

2) 变更请求 2

修复脚本为 ModifyPw,对应变更类型 4(删除界面元素),因删除"修改密码"链接,使得实现修改密码功能的其他界面元素不可达,所以提示用户 WebEdit("oldPw") 不可达,该测试脚本失效。

3) 变更请求 3

修复脚本 Tiaobao_HK、Tiaobao_YS 和 Tiaobao_ZJ,对应变更类型 2(界面元素类型替换),下列代码以 Tiaobao_ZJ 为例:

```
将单选框替换为下拉列表,将
WebRadioGroup("tianbaoType").Select DataTable("ZJ", "tianbao")
操作替换为
Set options = WebList("xpath:=//select[@name='type']").Object.all.tags("option")
For i = 0 to options.length -1
    If options(i).value = DataTable("ZJ", "tianbao") Then
        WebList("xpath:=//select[@name='type']").Select options(i).text
    End If
Next
```

4) 变更请求 4

修复脚本为 Tiaobao_ZJ,删除其中对于第三志愿下拉列表的操作,对应变更类型 4(删除界面元素);以及加上对新增表单元素联系电话输入框的操作,对应变更类型 3(新增界面元素)。下面为具体修复脚本代码片段:

```
删除对第三志愿操作
WebList("xpath:=//select[@name='zz_zj_zy3']").Select "#4"
在保存按钮提交前加入对联系电话输入框的操作
set input = Page().Object.getElementById("telephone")
input.value = "13888888888"
WebButton("保存").Click
```

5) 变更请求 5

修复脚本为 Print,对应变更类型 4(新增界面元素),在"志愿信息查看与打印"链接所在页面和志愿查看打印页面间加入了一个跳转按钮"确认",更改了原有的业务流程,需要加上对新增按钮的操作。下面为具体修复脚本代码:

```
Link("[志愿信息查看与打印]").Click
WebButton("xpath:=//input[@name='okBtn']").Click
WebButton("打印").Click
```

最后我们在新版本的系统中运行修复后的测试脚本,正确地通过回放,测试脚本的修复情况,符合预期。

4 结　语

因为客户对业务需求的变更请求,GUI 应用程序或者 Web 系统通常会发生演化,界面也会发生演化,如删除、新增或者改变某些界面元素,这往往会导致原有的测试脚本失效。本文提出了一种基于需求追踪的测试脚本修复方法,在业务需求、界面元素和测试脚本建立起追踪性,当用户发起变更请求时,开发人员记录下影响的变更元素,通过追踪性可以分析出潜在可能受影响的测试脚本集。本文分析了几种常见的导致测试脚本失效的界面元素变更类型,并给出了相应的测试脚本修复规则,还实现了面向测试工具 QTP 测试脚本修复工具,可以读入需求追踪文件,变更请求文件,对每个变更元素影响的测试脚本进行修复。最后通过一个实际的 Web 系统的演化案例,证明了方案的可行性。

目前,本文中采用的方法都是静态的文件分析方法,且所支持的 Web 系统开发技术都是基于 J2EE,未来可以兼容更多的开发技术,如 ASP. NET[13] 等。同时为了减轻开发者记录页面元素变更的工作量,可以开发一个 eclipse 插件,辅助开发者。

参 考 文 献

[1] Bertolino A. Software testing research: Achievements, challenges, dreams[C]//2007 Future of Software Engineering. IEEE Computer Society, 2007:85-103.
[2] Memon A M, Soffa M L. Regression testing of GUIs[C]//ACM SIGSOFT Software Engineering Notes. ACM, 2003, 28(5):118-127.
[3] Berner S, Weber R, Keller R K. Observations and lessons learned from automated testing[C]//Proceedings of the 27th international conference on Software engineering. ACM, 2005:571-579.
[4] Grechanik M, Xie Q, Fu C. Experimental assessment of manual versus tool-based maintenance of GUI-directed test scripts[C]//Software Maintenance, 2009. ICSM 2009. IEEE International Conference on. IEEE, 2009:9-18.
[5] Grechanik M, Xie Q, Fu C. Maintaining and evolving GUI-directed test scripts[C]//Software Engineering, 2009. ICSE 2009. IEEE 31st International Conference on. IEEE, 2009:408-418.
[6] Choudhary S R, Zhao D, Versee H, et al. Water: Web application test repair[C]//Proceedings of the First International Workshop on End-to-End Test Script Engineering. ACM, 2011:24-29.
[7] Wilde E, Lowe D. XPath, XLink, XPointer, and XML: A practical guide to Web hyperlinking and transclusion[M]. Addison-Wesley Longman Publishing Co., Inc., 2002.
[8] Ramesh B, Stubbs C, Powers T, et al. Requirements traceability: Theory and practice[J]. Annals of software engineering, 1997, 3(1): 397-415.
[9] Salem A M. Improving software quality through requirements traceability models[C]//Computer Systems and Applications, 2006. IEEE International Conference on. IEEE, 2006:1159-1162.
[10] Peraldi-Frati M A, Albinet A. Requirement traceability in safety critical systems[C]//Proceedings of the 1st Workshop on Critical Automotive applications: Robustness & Safety. ACM, 2010:11-14.
[11] Singh I, Brydon S, Murray G, et al. Designing Web Services with the J2EE 1.4 Platform: JAX-RPC, XML Services, and Clients[M]. Pearson Education, 2004.
[12] Marini J. Document Object Model[M]. McGraw-Hill, Inc., 2002.
[13] ASP. Net 动态网站编程指南[M]. 机械工业出版社,2001.