中国农业大学学报 2007,12(5):79-85
Journal of China Agricultural University

基于模糊系数规划的模糊支持向量分类机

杨志民[1] 田英杰[2]
(1.浙江工业大学 之江学院,杭州 310024; 2.中国科学院 数据技术与知识经济研究中心,北京 100080)

摘 要 本模糊支持向量分类机的构建特点是,训练点输出的类型和最终的模糊分类函数的函数值均为反映其模糊类别的实数。以模糊系数规划为基础,将模糊分类问题转化为求解模糊系数规划问题,求出模糊系数规划的 γ-最优规划,据此给出模糊支持向量分类机(算法);用 2 个例子说明该算法的合理性;最后给出模糊支持向量分类机中最佳阈值的确定方法。

关键词 机器学习; 模糊支持向量机分类; 模糊系数规划; 三角模糊数
中图分类号 TP 13; O 159 **文章编号** 1007-4333(2007)05-0079-07 **文献标识码** A

Fuzzy support vector classifier based on fuzzy coefficient programming

Yang Zhimin[1], Tian Yingjie[2]
(1. Zhijiang College, Zhejiang University of Technology, Hangzhou 310024, China;
2. Research Center of Data Technology and Knowledge Economy, Chinese Academy of Science, Beijing 100080, China)

Abstract This paper is concerned with a fuzzy support vector classifier in which the type of the figures for both the output of the training point and the value of the final fuzzy classification function is real number. The characteristic of real number is propitious to indicate the class of the figures as either positive or negative one. First, a fuzzy classification problem was formulated as a fuzzy coefficient programming problem. Then this programming was transformed into its optimal programming. As a result, we proposed a fuzzy support vector classification algorithm. In order to show its rationality, two examples were presented. In addition, we also proposed a strategy to decide the optimal threshold value in our algorithm.

Key words machine learning; fuzzy support vector classification; fuzzy coefficient programming; triangle fuzzy number

支持向量机是 Vapnik 等提出的一类新型机器学习方法[1-4],由于其出色的学习性能,该技术已成为机器学习界的研究热点,并在很多领域得到了成功的应用;但是,作为一种尚未成熟的新技术,有待于进一步完善。例如:处理带有模糊信息[5-6]的问题时,训练集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$ 中训练点的输出 $y_j(j=1,2,\cdots,l)$ 含有模糊隶属度。文献[7]中提出的 FSVM 方法只是对支持向量分类机(算法)中二次规划的惩罚参数添加了模糊隶属度,没有从算法的数学本质上建立模糊支持向量分类机,最终得到的分类函数的函数值(测试点的输出)还是确定值 1 或 - 1(正类或负类)。文献[8]和[9]分别建立了新模糊支持向量机(NFSVM)和加权支持向量机(PSVM),在 NFSVM 中用 fuzzy membership 和 fuzzy margin 反映训练点的模糊信息,在 PSVM 中用 possibilistic membership 和 possibilistic distance 表示训练点中的模糊信息,从而将模糊分类问题转化为求解二次规划[10-11]问题;但是最终得到的分类函数值(测试点的输出)还是确定值 1 或 - 1(正类或负类)。为此,笔者提出基于模糊系数规划的模糊支持向量分类机方法,其构建特点是使训练点输出的类型和最终的模糊分类函数的函数值均为反映其模糊类别的实数。

1 模糊支持向量分类机

在模糊分类中,样本点的输出是以隶属度形式

收稿日期: 2006-11-21
基金项目: 国家自然科学基金资助项目(10601064);浙江省自然科学基金资助项目(Y606082)
作者简介: 杨志民,教授,博士,主要从事支持向量机、不确定信息处理研究,E-mail: yzm9966@126.com

给出的,即样本点的输入 ——正类或负类的隶属度分别是 δ^+ 或 δ^- (δ^+, δ^- $\in[0.5,1.0]$)。为了表示方便,引进对应关系

$$\delta=\begin{cases}\delta^+ & \text{当输入 } x \text{ 为正类的隶属度为 } \delta^+ \text{ 时}\\ -\delta^- & \text{当输入 } x \text{ 为负类的隶属度为 } \delta^- \text{ 时}\end{cases} \tag{1}$$

式中 $\delta\in[-1.0,-0.5]\cup[0.5,1.0]$。这样可用 (x,δ) 表示样本点的输入 x 和它对应的输出,因此训练集形式为

$$S=\{(x_1,\delta_1),(x_2,\delta_2),\cdots,(x_l,\delta_l)\} \tag{2}$$

式中:$x_j\in R^n$ (R^n 为 n 维实空间);δ_j 与 δ 含义相同,$j=1,2,\cdots,l$。

为研究问题需要,给出如下转换规则,将式(2)中的 δ_j 转换为特殊的三角模糊数

$$\gamma_j=(r_{j1},r_{j2},r_{j3})=\begin{cases}\left(\dfrac{2\delta_j^2+\delta_j-2}{\delta_j},2\delta_j-1,\dfrac{2\delta_j^2-3\delta_j+2}{\delta_j}\right) & 0.5\le\delta_j\le1.0\\ \left(\dfrac{2\delta_j^2+3\delta_j+2}{\delta_j},2\delta_j+1,\dfrac{2\delta_j^2-\delta_j-2}{\delta_j}\right) & -1.0\le\delta_j\le-0.5\end{cases} \tag{3}$$

因此训练集式(2)可表示为

$$\overline{S}=\{(x_1,\gamma_1),(x_2,\gamma_2),\cdots,(x_l,\gamma_l)\} \tag{4}$$

式中 γ_j 为形如式(3)的三角模糊数,$j=1,2,\cdots,l$。

定义 1 式(2)中的 (x_j,δ_j) 和式(4)中的 (x_j,γ_j) 称为模糊训练点,$j=1,2,\cdots,l$;而 S 和 $\overline{S}$ 称为模糊训练集。

定义 2 在式(2)和(3)中,若 $\delta_j\in(0.5,1.0]$,则称其对应的模糊训练点为模糊正类点;若 $\delta_j\in[-1.0,-0.5)$,则称其对应的模糊训练点为模糊负类点。

说明:1)为简单起见,在此忽略 $\delta_j=0.5$ 或 $\delta_j=-0.5$ 的情形,因为此时对应的三角模糊数 $\gamma_j=(-2,0,2)$ 不提供正负类信息。

2)在模糊训练集式(2)和(4)中,将模糊正类点输出看作 1,将模糊负类点输出看作 -1,得到普通训练集 T。若 T 线性可分,则模糊训练集 S 和 $\overline{S}$ 线性可分。

为研究方便,将模糊训练集式(2)和(4)中的模糊训练点重新排序,将模糊正类点排在前面,将模糊负类点排在后面,得到 2 个模糊训练集

$$S=\{(x_1,\delta_1),(x_2,\delta_2),\cdots,(x_p,\delta_p),(x_{p+1},\delta_{p+1}),\cdots,(x_l,\delta_l)\} \tag{5}$$

和

$$\overline{S}=\{(x_1,\gamma_1),(x_2,\gamma_2),\cdots,(x_p,\gamma_p),(x_{p+1},\gamma_{p+1}),\cdots,(x_l,\gamma_l)\} \tag{6}$$

设 $t=1,2,\cdots,p$, $i=p+1,p+2,\cdots,l$,则式中 (x_t,δ_t) 和 (x_t,γ_t) 为模糊正类点,(x_i,δ_i) 和 (x_i,γ_i) 为模糊负类点。

设给定形如式(5)的线性可分问题的模糊训练集。首先,将模糊训练集式(5)按转换规则式(3)转换为模糊训练集式(6)。此时模糊分类问题转化为求解以 $(w,b)^{\mathrm{T}}$ 决策变量的模糊系数规划问题[12]

$$\begin{cases}\min\limits_{w,b,\xi}\dfrac{1}{2}\|w\|^2\\ \text{s.t. }\gamma_j((w\cdot x_j)+b)\ge1\quad j=1,2,\cdots,l\end{cases} \tag{7}$$

以下定理给出了模糊系数规划式(7)的 γ- 最优规划。

定理 1 对于给定阈值 $\gamma(0\le\gamma\le1)$,模糊系数规划式(7)的 γ- 最优规划为二次规划

$$\begin{cases}\min\limits_{w,b}\dfrac{1}{2}\|w\|^2\\ \text{s.t. }((1-\gamma)r_{t3}+\gamma r_{t2})((w\cdot x_t)+b)\ge1\\ \quad t=1,2,\cdots,p\\ ((1-\gamma)r_{i1}+\gamma r_{i2})((w\cdot x_i)+b)\ge1\\ \quad i=p+1,p+2,\cdots,l\end{cases} \tag{8}$$

式中:$(1-\gamma)r_{t3}+\gamma r_{t2}$ 为模糊正类点输出 γ_t 的 γ- 水平截集(闭区间)右端点,$(1-\gamma)r_{i1}+\gamma r_{i2}$ 为模糊负类点输出 γ_i 的 γ- 水平截集(闭区间)左端点。

式(8)为凸二次规划,所以克服了局部极小点的困难,因此有:

定理 2 定理 1 中的二次规划式(8)的最优解存在。

下面求二次规划式(8)的对偶规划。

定理 3 二次规划式(8)的对偶规划为

$$\begin{cases}\min\limits_{\beta,\alpha}\dfrac{1}{2}(A+2B+C)-\left(\sum\limits_{t=1}^{p}\beta_t+\sum\limits_{i=p+1}^{l}\alpha_i\right)\\ \text{s.t. }\sum\limits_{t=1}^{p}\beta_t((1-\gamma)r_{t3}+\gamma r_{t2})+\\ \quad\sum\limits_{i=p+1}^{l}\alpha_i((1-\gamma)r_{i1}+\gamma r_{i2})=0\\ \beta_t\ge0\quad t=1,2,\cdots,p\\ \alpha_i\ge0\quad i=p+1,p+2,\cdots,l\end{cases} \tag{9}$$

式中

$$A = \sum_{t=1}^{p}\sum_{s=1}^{p}\beta_t\beta_s((1-\gamma)\,r_{t3}+\gamma r_{t2})\cdot((1-\gamma)\,r_{s3}+\gamma r_{s2})(x_t\cdot x_s)$$

$$B = \sum_{t=1}^{p}\sum_{i=p+1}^{l}\beta_t\alpha_i((1-\gamma)\,r_{t3}+\gamma r_{t2})\cdot((1-\gamma)\,r_{i1}+\gamma r_{i2})(x_t\cdot x_i)$$

$$C = \sum_{i=p+1}^{l}\sum_{q=p+1}^{l}\alpha_i\alpha_q((1-\gamma)\,r_{i1}+\gamma r_{i2})\cdot((1-\gamma)\,r_{q1}+\gamma r_{i2})(x_i\cdot x_q)$$

$\beta=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$, $\alpha=(\alpha_{p+1},\alpha_{p+2},\cdots,\alpha_l)^{\mathrm{T}}$, $(\beta,\alpha)^{\mathrm{T}}$ 为决策变量。

证明请参阅文献[10]。

规划式(9)为一个凸二次规划,解得其最优解 $(\beta^*,\alpha^*)^{\mathrm{T}}=(\beta_1^*,\beta_2^*,\cdots,\beta_p^*,\alpha_{p+1}^*,\alpha_{p+2}^*,\cdots,\alpha_l^*)^{\mathrm{T}}$。可以证明,模糊系数规划式(7)的 γ- 最优解为 $(w^*,b^*)^{\mathrm{T}}$[10-11],其中

$$w^* = \sum_{t=1}^{p}\beta_t^*((1-\gamma)\,r_{t3}+\gamma r_{t2})x_t + \sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)\,r_{i1}+\gamma r_{i2})x_i$$

$$b^* = ((1-\gamma)\,r_{s3}+\gamma r_{s2}) - \left(\sum_{t=1}^{p}\beta_t^*((1-\gamma)\,r_{t3}+\gamma r_{t2})(x_t\cdot x_s) + \sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)\,r_{i1}+\gamma r_{i2})(x_i\cdot x_s)\right)\quad s\in\{s\mid\beta_s^*>0\}$$

或者

$$b^* = ((1-\gamma)\,r_{qi}+\gamma r_{q2}) - \left(\sum_{t=1}^{p}\beta_t^*((1-\gamma)\,r_{t3}+\gamma r_{t2})(x_t\cdot x_q) + \sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)\,r_{i1}+\gamma r_{i2})(x_i\cdot x_q)\right)\quad q\in\{q\mid\alpha_q^*>0\}$$

得到确定性最优分类超平面[10]

$$(w^*\cdot x)+b^*=0\quad x\in R^n$$

从而得到函数

$$g(x)=(w^*\cdot x)+b^* \tag{10}$$

因为在模糊支持向量分类机中,训练点与测试点的形式与隶属度规律要保持一致性,所以利用支持向量回归机构建模糊最优分类函数(函数值形如模糊训练集式(5)中的 δ_j, $j=1,2,\cdots,p,p+1,\cdots,l$)

$$\delta=\delta(g(x))=\begin{cases}\phi_+(g(x)) & 0<g(x)\le\phi_+^{-1}(1)\\ 1 & g(x)>\phi_+^{-1}(1)\\ -\phi_-(g(x)) & \phi_-^{-1}\le g(x)<0\\ -1 & g(x)<\phi_-^{-1}(1)\end{cases} \tag{11}$$

设 $g(x)=u$,则 $\phi_+(g(x))=\phi_+(u)$ 为由 ε- 支持向量回归机得到的回归函数(关于 u 的单调增函数)。此 ε- 支持向量回归机构造方法如下:

1) 构造回归问题的训练集

$$\{(g(x_1),\delta_1),(g(x_2),\delta_2),\cdots,(g(x_p),\delta_p)\} \tag{12}$$

2) 以式(12)为训练集,选择适当的 $\varepsilon>0$、惩罚参数 $C>0$,选择核函数为线性核,构造 ε- 支持向量回归机。

同理,$\phi_-(u)$ 为由 ε- 支持向量回归机得到的回归函数(关于 u 的单调减函数)。此 ε- 支持向量回归机构造方法如下:

1) 构造回归问题的训练集

$$\{(g(x_{p+1}),-\delta_{p+1}),(g(x_{p+2}),-\delta_{p+2}),\cdots,(g(x_l),-\delta_l)\} \tag{13}$$

2) 以式(13)为训练集,选择与构造 $\phi_+(u)$ 相同的 ε 和 C,选择核函数为线性核,构造 ε- 支持向量回归机。

$\phi_+^{-1}(1)$ 和 $\phi_-^{-1}(1)$ 分别为函数 $\phi_+(u)$ 和 $\phi_-(u)$ 的反函数在 1 处的函数值。

模糊最优分类函数是用 $\delta=\delta^+$ 和 $\delta=-\delta^-$ 来表示测试点的输入 x 为正类(或负类)的隶属度的,而样本点为正类(或负类)隶属度的一般规律是:对于模糊正类点,若 $g(x)$ 大,则相应的隶属度大;对于模糊负类点,若 $g(x)$ 小,则相应的隶属度大。因此利用回归函数 $\phi_+(u)$ 和 $\phi_-(u)$ 构造模糊最优分类函数 $\delta=\delta(g(x))$ 具有较大的合理性。

任给一测试点输入 $x'\in R^n$,代入模糊最优分类函数式(11)得到 $\delta(x)$,即为测试点的输出。它可以客观地反映测试点 $(x',\delta(x))$ 的模糊分类情况(即显示测试点的输入 x 为正类或负类的隶属度)。

通过以上分析,得到

算法 1 (线性可分模糊支持向量分类机)

1) 给定模糊训练集式(5),按转换规则式(3)将式(5)转换为模糊训练集式(6)。

2) 选择适当的阈值 $\gamma(0\le\gamma\le1)$,构造二次规划式(9)。

3) 求解二次规划式(9),得最优解$(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*)^{\mathrm{T}}=(\beta_1^*,\beta_2^*,\cdots,\beta_p^*,\alpha_{p+1}^*,\alpha_{p+2}^*,\cdots,\alpha_l^*)^{\mathrm{T}}$。

4) 计算 $w^*=\sum_{t=1}^{p}\beta_t^*((1-\gamma)r_{t3}+\gamma r_{t2})x_t+\sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)r_{i1}+\gamma r_{i2})x_i$。选择 $\boldsymbol{\beta}^*$ 的正分量 β_s^* 或 $\boldsymbol{\alpha}^*$ 的正分量 α_q^*,据此计算

$$b^*=((1-\gamma)r_{s3}+\gamma r_{s2})-\left(\sum_{t=1}^{p}\beta_t^*((1-\gamma)r_{t3}+\gamma r_{t2})\cdot(x_t\cdot x_s)+\sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)r_{i1}+\gamma r_{i2})(x_i\cdot x_s)\right)$$

或

$$b^*=((1-\gamma)r_{qi}+\gamma r_{q2})-\left(\sum_{t=1}^{p}\beta_t^*((1-\gamma)r_{t3}+\gamma r_{t2})\cdot(x_t\cdot x_q)+\sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)r_{i1}+\gamma r_{i2})(x_i\cdot x_q)\right)$$

5) 构造函数式(10)。

6) 分别以式(12)和(13)为训练集,构造 ε- 支持向量回归机(选择适当 ε、惩罚参数 C,选择核函数为线性核),得到回归函数 $\phi_+(u)$ 和 $\phi_-(u)$。

7) 构造模糊最优分类函数式(11)。

对于形如式(5)的近似线性可分问题的模糊训练集,首先,将模糊训练集式(5)按转换规则式(3)转换为模糊训练集式(6),此时模糊分类问题转化为求解如下以$(w,b,\boldsymbol{\xi})^{\mathrm{T}}$为决策变量的模糊系数规划问题:

$$\begin{cases}\min\limits_{w,b,\xi}\dfrac{1}{2}\|w\|^2+C\sum\limits_{j=1}^{l}\xi_j\\ \text{s.t. } \tilde{y}_j((w\cdot x_j)+b)+\xi_j\geq 1\\ \xi_j\geq 0\end{cases}\qquad j=1,2,\cdots,l \tag{14}$$

式中:$C>0$ 是惩罚参数;$\boldsymbol{\xi}=(\xi_1,\xi_2,\cdots,\xi_l)^{\mathrm{T}}$。

模糊系数规划式(14)的 γ- 最优规划和对偶规划,以及近似线性可分问题的模糊最优分类函数等与线性可分问题类似,在此从略。

对于非线性问题,模糊支持向量分类的构造方法与线性问题类似,因此得到

算法 2 (非线性模糊支持向量分类机)

1) 给定模糊训练集式(5),按转换规则式(3)将式(5)转换为模糊训练集式(6)。

2) 选择适当的阈值 $\gamma(0\leq\gamma\leq 1)$、惩罚参数 C,以及适当的核函数$(K(x,x'))$,构造二次规划:

$$\begin{cases}\min\limits_{\boldsymbol{\beta},\boldsymbol{\alpha}}\dfrac{1}{2}(A_K+2B_K+C_K)-\left(\sum\limits_{t=1}^{p}\beta_t+\sum\limits_{i=p+1}^{l}\alpha_i\right)\\ \text{s.t. }\sum\limits_{t=1}^{p}\beta_t((1-\gamma)r_{t3}+\gamma r_{t2})+\sum\limits_{i=p+1}^{l}\alpha_i((1-\gamma)r_{i1}+\gamma r_{i2})=0\\ 0\leq\beta_t\leq C\quad t=1,2,\cdots,p\\ 0\leq\alpha_i\leq C\quad i=p+1,p+2,\cdots,l\end{cases} \tag{15}$$

式中:

$$A_K=\sum_{t=1}^{p}\sum_{s=1}^{p}\beta_t\beta_s((1-\gamma)r_{t3}+\gamma r_{t2})((1-\gamma)r_{s3}+\gamma r_{s2})K(x_t,x_s)$$

$$B_K=\sum_{t=1}^{p}\sum_{i=p+1}^{l}\beta_t\alpha_i((1-\gamma)r_{t3}+\gamma r_{t2})((1-\gamma)r_{i1}+\gamma r_{i2})K(x_t,x_i)$$

$$C_K=\sum_{i=p+1}^{l}\sum_{q=p+1}^{l}\alpha_i\alpha_q((1-\gamma)r_{i1}+\gamma r_{i2})((1-\gamma)r_{q1}+\gamma r_{i2})K(x_i,x_q)$$

$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}\in R_+^p$,$\boldsymbol{\alpha}=(\alpha_{p+1},\alpha_{p+2},\cdots,\alpha_l)^{\mathrm{T}}\in R_+^{1-p}$,$(\boldsymbol{\beta},\boldsymbol{\alpha})^{\mathrm{T}}$ 为决策变量。

3) 求解二次规划式(15),得最优解

$$(\boldsymbol{\beta}^*,\boldsymbol{\alpha}^*)^{\mathrm{T}}=(\beta_1^*,\beta_2^*,\cdots,\beta_p^*,\alpha_{p+1}^*,\alpha_{p+2}^*,\cdots,\alpha_l^*)^{\mathrm{T}}$$

4) 选择 $\boldsymbol{\beta}^*$ 的正分量 $0<\beta_s^*<C$ 或 $\boldsymbol{\alpha}^*$ 的正分量 $0\leq\alpha_q^*<C$,据此计算

$$b^*=((1-\gamma)r_{s3}+\gamma r_{s2})-\left(\sum_{t=1}^{p}\beta_t^*((1-\gamma)r_{t3}+\gamma r_{t2})\cdot K(x_t,x_s)+\sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)r_{i1}+\gamma r_{i2})K(x_i,x_s)\right)$$

或

$$b^*=((1-\gamma)r_{qi}+\gamma r_{q2})-\left(\sum_{t=1}^{p}\beta_t^*((1-\gamma)r_{t3}+\gamma r_{t2})K(x_t,x_q)+\sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)r_{i1}+\gamma r_{i2})K(x_i,x_1)\right)$$

5) 构造函数

$$g(x)=\sum_{t=1}^{p}\beta_t^*((1-\gamma)r_{t3}+\gamma r_{t2})K(x,x_t)+\sum_{i=p+1}^{l}\alpha_i^*((1-\gamma)r_{i1}+\gamma r_{i2})K(x,x_i)+b^*$$

6) 分别以$\{(g(x_1),\delta_1),(g(x_2),\delta_2),\cdots,(g(x_p),\delta_p)\}$和$\{(g(x_{p+1}),-\delta_{p+1}),(g(x_{p+2}),-\delta_{p+2}),\cdots,(g(x_l),-\delta_l)\}$为训练集,构造 ε- 支

持向量回归机(选择适当 ε、惩罚参数 C,选择核函数为线性核),得到回归函数 $\phi_+(u)$ 和 $\phi_-(u)$。

7) 构造模糊最优分类函数

$$\delta=\delta(x)=\begin{cases}\phi_+(g(x)) & 0<g(x)\leq\phi_+^{-1}(1)\\ 1 & g(x)>\phi_+^{-1}(1)\\ -\phi_-(g(x)) & \phi_-^{-1}\leq g(x)<0\\ -1 & g(x)<\phi_-^{-1}(1)\end{cases}.$$

若模糊训练集式(5)中所有模糊训练点输出 δ_j($j=1,2,\cdots,l$)全为 1 或 -1,则模糊训练集退化为普通训练集,模糊支持向量分类机变为支持向量分类机.

2　数值试验与应用实例

以一个简单的例子说明所建立算法的合理性。

设模糊训练集为

$$S=\{(x_1,\delta_1),(x_2,\delta_2),(x_3,\delta_3)\}$$

式中: $x_1=3.00$, $\delta_1=1.00$; $x_3=-1.00$, $\delta_3=-1.00$; $x_2=1.00$, $\delta_2\in[-1.00,-0.50]\cup[0.50,1.00]$,即让 δ_2 在区间 $[-1.00,-0.50]\cup[0.50,1.00]$ 中变化。

取阈值 $\gamma=0.20$,根据算法 1(线性可分模糊支持向量分类机)可得

$$w^*=\begin{cases}\frac{1}{2}\left(1+\frac{1}{\Delta^+}\right) & \delta_2\in[0.50,1.00]\\ \frac{1}{2}\left(1-\frac{1}{\Delta^-}\right) & \delta_2\in[-1.00,-0.50]\end{cases}$$

$$b^*=\begin{cases}\frac{1}{2}\left(1+\frac{1}{\Delta^+}\right)-1 & \delta_2\in[0.50,1.00]\\ 1-\frac{3}{2}\left(1-\frac{1}{\Delta^-}\right) & \delta_2\in[-1.00,-0.50]\end{cases}$$

因此确定性最优分类超平面为

$$x=\frac{b^*}{w^*}=\begin{cases}\frac{\Delta^+-1}{\Delta^++1} & \delta_2\in[0.50,1.00]\\ \frac{\Delta^--3}{\Delta^--1} & \delta_2\in[-1.00,-0.50]\end{cases}$$

式中:

$$\Delta^+=\frac{\frac{4}{5}(2\delta_2^2-3\delta_2+2)}{\delta_2}+\frac{1}{5}(2\delta_2-1)$$

$$\Delta^-=\frac{\frac{4}{5}(2\delta_2^2+3\delta_2+2)}{\delta_2}+\frac{1}{5}(2\delta_2+1)$$

取 $\delta_2=1.00$,得确定性最优分类超平面 $x=0$; $\delta_2=0.90$,得 $x=0.11$; $\delta_2=0.80$,得 $x=0.23$; $\delta_2=0.70$,得 $x=0.36$; $\delta_2=0.60$,得 $x=0.62$; $\delta_2=0.50$,得 $x=1.00$; $\delta_2=-0.50$,得 $x=1.00$; $\delta_2=-0.60$,得 $x=1.36$; $\delta_2=-0.70$,得 $x=1.61$; $\delta_2=-0.80$,得 $x=1.76$; $\delta_2=-0.90$,得 $x=1.88$; $\delta_2=-1.00$,得 $x=2.00$(图 1)。

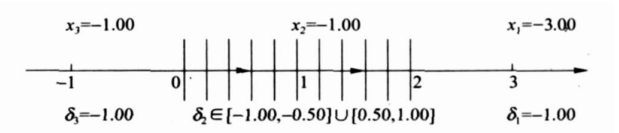

图 1　模糊分类图例

Fig. 1　Schematic diagram of fuzzy classification

当模糊训练点 (x_2,δ_2) 的输出由 $\delta_2=1.00$ 逐渐变化为 $\delta_2=-1.00$ 时,确定性最优分类超平面从 $x=0$ 逐渐向 $x=2.00$ 移动。所得结果与直观判断相吻合。

为了体现本算法的应用价值,给出一个具体应用实例。将模糊支持向量分类机应用于河北省邯郸市粮食安全预警,得出基于模糊支持向量分类机(算法)的邯郸市粮食安全预警方法。

以财政支农增长指数(X_1)、财政农业基本建设投资增长指数(X_2)、科技 3 项费用指数(X_3)、农机总动力指数(X_4)、有效灌溉面积指数(X_5)、排灌机械动力指数(X_6)、粮食价格指数(X_7)7 项指标作为警兆指标,以 1988—1997 年的数据作为训练样本集,以 1998、1999、2000、2001、2002 年作为测试样本,进行预警试验。咨询了粮食专家,就 1988—1997 年的粮食安全历史警度进行分析。由于历史资料不够全面,历史数据不够准确,给专家判断分类(有警,无警)带来一定影响;因此专家只能给出历史警度的模糊判别,将其转化为"有警隶属度"(表 1)。根据模糊支持向量分类机(算法)得到对邯郸市 1998、1999、2000、2001、2002 年的粮食安全预警结果(表 2),同时利用支持向量机(算法)和模糊聚类方法,采用同样数据对邯郸市 1998、1999、2000、2001、2002 年的粮食安全进行预警,结果见表 3。

粮食专家对上述 3 种方法得到的预警结果进行评价,得出结论:

1) 利用模糊支持向量分类机(算法)得到的预警结果符合实际,客观地反映了邯郸市 1998—2002 年粮食安全的实际情况;

2) 利用支持向量机(算法)得到的预警结果基本符合实际,但反映邯郸市 1998—2002 年粮食安全的实际情况不太客观(缺乏实际问题的中间过渡);

3) 利用模糊聚类方法得到的预警结果与实际情况偏差较大。

由此看出,模糊支持向量分类机(算法)处理模糊分类问题有较大的优越性。

表 1 邯郸市粮食安全历史警度

Table 1 History warning degree of grain safe in Handan City

年份	X_1	X_2	X_3	X_4	X_5	X_6	X_7	有警隶属度
1988	9.376 0	-15.253 0	4.824 6	7.001 9	-0.061 0	6.308 0	6.131 2	1.000 0
1989	24.230 0	27.653 1	3.765 7	5.614 3	1.219 1	6.971 0	10.732 0	0.369 0
1990	15.760 0	31.733 8	25.413 0	2.382 1	84.551 0	4.412 0	-2.560 0	0.291 0
1991	12.910 0	13.161 4	-5.778 0	2.771 8	0.873 9	2.945 0	0.643 8	0.893 0
1992	8.185 0	12.589 6	2.388 1	3.119 8	1.616 2	-0.163 0	1.361 0	0.598 0
1993	17.130 0	11.663 7	0	4.996 1	0.283 6	4.976 2	-3.362 0	0.697 0
1994	21.110 0	12.621 6	0.112 4	6.724 4	0.063 1	1.727 0	3.112 0	0.923 0
1995	7.881 0	2.813 3	0.000 1	6.113 2	1.708 9	2.899 0	5.012 6	0.488 0
1996	9.417 0	12.910 7	10.931 0	8.796 5	1.712 2	7.717 8	7.998 9	0.235 0
1997	50.670 0	88.564 0	56.987 0	7.967 4	2.631 0	4.823 1	7.225 1	0

注:若隶属度为 1.000 0,则确定有警;若隶属度为 0,则确定无警;若隶属度为 $\alpha(0<\alpha<1)$,则隶属于有警的程度为 α。

表 2 邯郸市粮食安全预警结果 Ⅰ(模糊支持向量分类机方法)

Table 2 Result Ⅰ of grain safe early-warning in Handan City (by the method of fuzzy support vector classification)

年份	X_1	X_2	X_3	X_4	X_5	X_6	X_7	有警隶属度
1998	49.320 0	37.234 0	15.331 0	20.012 0	7.554 6	5.886 1	9.687 7	0.120 0
1999	9.376 0	-15.253 0	4.824 6	7.001 9	-0.061 0	6.308 1	6.112 5	0.850 0
2000	6.254 0	6.701 2	-19.552 0	8.254 0	-8.658 0	4.075 8	-1.729 0	0.980 0
2001	41.120 0	33.123 0	10.289 0	15.876 2	8.442 3	4.556 4	0.235 2	0.410 0
2002	46.350 0	35.117 0	16.524 0	19.331 0	2.115 0	4.362 4	8.772 2	0.230 0

表 3 邯郸市粮食安全预警结果 Ⅱ (支持向量机和模糊聚类方法)

Table 3 Result Ⅱ of grain safe early-warning in Handan City (by the support vector machine methods and the fuzzy clustering technique)

年份	支持向量机预警结果(将模糊训练点近似为清晰)	模糊聚类方法预警结果
1998	无警	无警
1999	有警	有警
2000	有警	有警
2001	无警	有警
2002	无警	无警

3 最佳阈值的确定

在支持向量分类机中有惩罚参数 C 和核函数 $K(x, x')$,而在模糊支持向量分类机中除这 2 个参数之外还有参数阈值 $\gamma(0 \leq \gamma \leq 1)$。

设模糊训练集如式(5),则确定最佳阈值 γ 的具体方法如下:

1) 在 l 个模糊训练点中任取一个 (x_k, δ_k) 作为测试点,其余 $l-1$ 个作为训练点。

2) 将区间 $[0,1]$ 进行 a 等分,得分点 $a_j = j/a$, $j = 1, 2, \cdots, a$。

3) 取隶属度 $\gamma = a_j$,构造模糊支持向量分类机,得模糊最优分类函数 $\delta(x)$。将测试点的输入 x_k 代入 $\delta(x)$ 中,得 $\delta_k = \delta(x_k)$。令 $\delta_{aj,k} = |\delta_k - \delta_k|$,称 $\delta_{a_j,k}$ 为 $\gamma = a_j (j = 1, 2, \cdots, a)$ 时模糊支持向量分类机在 (x_k, δ_k) 处的误差。

4) 当取遍集合 $\{1, 2, \cdots, l\}$ 时,对于阈值 $\gamma = a_j$ $(j = 1, 2, \cdots, a)$,得到 l 个误差 $\delta_{a_j,1}, \delta_{a_j,2}, \cdots, \delta_{a_j,l}$。

令 $\delta_{a_j}=\sum_{k=1}^{l}\delta_{a_j,k}$,称 δ_{a_j}为 $\gamma=a_j(j=1,2,\cdots,a)$模糊支持向量分类机的总误差。

5)选择 ρ,使 $\delta_\rho=\min_{j=1,2,\cdots,a}\{\delta_{a_j}\}$,则取 $\gamma=\rho$。ρ称为最佳阈值。

4 讨　论

本研究训练点中含有完整模糊信息(即训练点的输入为正类与负类的隶属度之和为 1)的模糊分类问题。本文算法中,训练点的输出为三角模糊数,建立的模糊最优分类函数,其函数值也为三角模糊数。本文算法与文献[7-9]3 种算法的相同点为:都是利用最优化方法解决训练点中含有模糊信息的模糊分类问题。区别为:本文算法以模糊优化方法为工具处理训练点中含有模糊信息的模糊分类问题,使训练点中的模糊信息自然反映在模糊规划中,得出的模糊最优分类函数仍含有模糊性,即任给一测试点输入,代入模糊最优分类函数中,得到的输出为三角模糊数,从而使测试点与训练点形式匹配,逻辑一致;而其他 3 种方法以普通优化方法为工具处理训练点中含有模糊信息的模糊分类问题,因此最终得到的分类函数为确定函数(不含有模糊性),从而使测试点的输出为确定的正类或负类,与训练点的输出(模糊隶属度)形式不匹配,这样在利用训练点作为测试点做 LOO 误差估计[10]进行参数选择时就出现了错误。

5 结　论

本研究在支持向量机和模糊系数规划基础上,建立了模糊支持向量分类机(算法),并且将其应用于邯郸市粮食安全预警,得出基于模糊支持向量分类机的邯郸市粮食安全预警方法。1988—2002 年邯郸市粮食安全预警结果表明,模糊支持向量分类机方法优于支持向量机方法和模糊聚类方法。

参 考 文 献

[1] Vapnik V N. Statistical Learning Theory[M]. New York: Wiley, 1998

[2] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995

[3] Cristianini N, Shawe-Taylor J. Introduction to Support Vector Machines[M]. Cambridge: Cambridge University Press, 2000

[4] Mangasarian O L. Multi-surface method of pattern separation[J]. IEEE Transaction on Information Theory, 1968(14): 801-807

[5] Zadeh L A. Fuzzy sets[J]. Information and Control, 1965(8): 338-353

[6] 张文修. 模糊数学基础[M]. 西安: 西安交通大学出版社, 1995

[7] Lin Chunfu, Wang Shengde. Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 2002(2): 464-471

[8] Tao Qing, Wang Jue. A new fuzzy support vector machine based on the weighted margin[J]. Neural Procession Letters, 2004(3): 139-150

[9] Lee Kiyoung, Kim Dae-Won, Lee Doheon, et al. Possibilistic support vector machines[J]. Pattern Recognition, 2005(38): 1325-1327

[10] 邓乃扬, 田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京: 科学出版社, 2004

[11] 邓乃扬, 诸梅芳. 最优化方法[M]. 沈阳: 辽宁教育出版社, 1987

[12] 曾庆宁. 模糊系数规划[J]. 模糊系统与数学, 2000(3): 99-105