

未知支持向量机

杨志民¹ 邵元海¹ 梁静²

摘要 提出一种处理样本中含有未知信息(一种不确定性信息)的支持向量机——未知支持向量机(Unascertained support vector machine, USVM)算法。首先,以未知数学为基础,将含有未知信息的分类问题转化为求解未知机会约束规划问题。然后,将其转化为与其等价的二次规划。据此给出未知支持向量机。理论分析和试验结果均表明,该算法是有效、可行的。

关键词 机器学习,未知支持向量机,未知信息,未知数

引用格式 杨志民,邵元海,梁静.未知支持向量机.自动化学报,2013,39(6):895–901

DOI 10.3724/SP.J.1004.2013.00895

Unascertained Support Vector Machine

YANG Zhi-Min¹ SHAO Yuan-Hai¹ LIANG Jing²

Abstract In this paper, we propose an unascertained support vector machine (USVM) for classification problem with unascertained information (a kind of uncertain information). First, based on unascertained mathematics, classification problems containing unascertained information are converted to solving unascertained chance constrained programming problems. Then, the unascertained chance constrained programming is converted to its equivalent quadratic programming. Based on these theories, the algorithm of a unascertained support vector machine is given. Theoretical analysis and experiments show that the proposed USVM is effective and feasible.

Key words Machine learning, unascertained support vector machine (USVM), unascertained information, unascertained number

Citation Yang Zhi-Min, Shao Yuan-Hai, Liang Jing. Unascertained support vector machine. *Acta Automatica Sinica*, 2013, 39(6): 895–901

研究信息,特别是研究不确定性信息的表达和处理是当今信息时代的一个重要问题。未知信息是不同于随机信息、模糊信息和粗糙信息的一种不确定性信息,它的不确定性不在于事物本身,而是由于决策者不能完全把握事物的真实状态和数量关系,造成的纯主观认识上的不确定性。在决策中就不能把其看成是确定的,而必须当作不确定事物对待。上世纪 90 年代中国工程院院士、哈尔滨工业大学王光远教授、河北工程大学刘开第教授等学者创立了“未知数学”^[1–3],用以表达和处理未知信息。20 多年来在广大学者的不懈努力下,“未知数学”从理论知识到实际应用都得到了长足的发展。尤其是“未知数学”在众多领域(如:数字滤波、企业投产

可行性分析、整体优化、水质分析、煤矿立井施工等)的应用^[4–6],更显示出其处理未知信息的优越性。

支持向量机是 Vapnik 等提出的一类新型机器学习方法^[7–9],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。但是支持向量机在处理某些问题时还存在许多局限,比如:客观世界存在大量的未知信息,如果支持向量机的样本中含有未知信息^[1–3, 5–6],那么标准支持向量机将无能为力。而且目前许多学者研究的模糊支持向量机^[10–18]是以模糊数学为基础处理样本中含有模糊信息的分类问题,对于含有未知信息分类问题无法很好地解决。基于以上因素我们在标准支持向量分类机基础上,以未知数学和未知规划为工具^[1–2, 4–5, 19–22],根据未知信息的特点,建立未知支持向量机。

未知支持向量机与模糊支持向量机的相同点为它们都是处理样本中含有不确定性信息的分类问题。而它们的区别在于样本中含有的不确定性信息不同(一个为未知信息,另一个为模糊信息),建立算法所用的工具不同(一个为未知规划,另一个为模糊规划),从而得到的算法不同(一个为处理样本中

收稿日期 2012-03-22 录用日期 2012-10-31

Manuscript received March 22, 2012; accepted October 31, 2012
国家自然科学基金(10926198, 11201426),浙江省自然科学基金(LQ12A01020)资助

Supported by National Natural Science Foundation of China (10926198, 11201426) and the Zhejiang Provincial Natural Science Foundation of China (LQ12A01020)

本文责任编辑 王聪

Recommended by Associate Editor WANG Cong

1. 浙江工业大学之江学院 杭州 310024 2. 陕西煤业化工集团黄陵矿业股份有限公司 黄陵 727307

1. Zhijiang College, Zhejiang University of Technology, Hangzhou 310024 2. Shaanxi Coal Chemical Group, Huangling Mining Industry, Huangling 727307

含有未确知信息的分类算法, 另一个为处理样本中含有模糊信息的分类算法). 对于样本中含有未确知信息的分类问题, 用未确知支持向量机 (算法) 能较较好地处理. 而若用模糊支持向量机 (算法) 进行处理, 由于理论算法上的差异, 就容易出现较大的误差, 在数值实验中我们将进一步用数据进行说明.

在文献 [20] 中, 我们提出未确知机会约束规划和构建未确知支持向量机的基本思路, 尚未形成算法. 在此基础上本文构建了未确知支持向量机完整算法, 进行了严格证明, 并且用数据实验显示了算法的有效性.

1 预备知识

对于未确知信息, 我们可用 “未确知数”^[1-2, 4-5] 和 “未确知规划”^[19-22] 进行表示或描述. 下面对未确知数和未确知规划做一简介, 为建立未确知支持向量机做准备.

定义 1. 对任意闭区间 $[a, b]$, $a = y_1 < y_2 < \cdots < y_n = b$, 若函数 $f(y)$ 满足

$$f(y) = \begin{cases} \alpha_i, & y = y_i, \quad i = 1, \cdots, n \\ 0, & y \neq y_i, \quad y \in [a, b] \end{cases} \quad (1)$$

且 $\sum_{i=1}^n \alpha_i = \alpha$, $0 < \alpha_i \leq 1$ ($i = 1, \cdots, n$), 则称 $[a, b]$ 和 $f(y)$ 构成一个 n 阶未确知数, 记作 $[[a, b], f(y)]$. 称 α 、 $[a, b]$ 和 $f(y)$ 分别为该未确知数的总可信度、取值区间和可信度分布密度函数. $\{y_i\}$ 称为该未确知数的可能值序列, $\{\alpha_i\}$ 称为该未确知数的可信度序列 ($i = 1, \cdots, n$).

定义 2. 设未确知数 $A = [[a, b], f(y)]$, 其中 $f(y)$ 满足式 (1) 且 $0 < \sum_{i=1}^n \alpha_i = \alpha \leq 1$, $0 < \alpha_i \leq 1$ ($i = 1, \cdots, n$). 若函数 $F(y)$ 满足

$$F(y) = \begin{cases} 0, & y < a \\ \vdots & \vdots \\ \sum_{i=1}^k \alpha_i, & y_k \leq y < y_{k+1}, \quad k = 1, \cdots, n-1 \\ \vdots & \vdots \\ \alpha, & y \geq b \end{cases} \quad (2)$$

则称闭区间 $[a, b]$ 与函数 $F(y)$ 构成分布函数型未确知数, 记作 $[[a, b], F(y)]$. 称 α 、 $[a, b]$ 和 $F(y)$ 分别为该分布函数型未确知数的总可信度、分布区间和可信度分布函数.

定义 3. 设二未确知数 $A = [[x_1, x_n], f(x)]$, $B = [[y_1, y_m], g(y)]$, 其中

$$f(x) = \begin{cases} \alpha_i, & x = x_i, \quad i = 1, \cdots, n \\ 0, & x \neq x_i, \quad x \in [x_1, x_n] \end{cases} \quad (3)$$

$$\sum_{i=1}^n \alpha_i = \alpha, \quad 0 < \alpha \leq 1, \quad i = 1, \cdots, n \quad (4)$$

$$g(y) = \begin{cases} \beta_j, & y = y_j, \quad j = 1, \cdots, m \\ 0, & y \neq y_j, \quad y \in [y_1, y_m] \end{cases} \quad (5)$$

$$\sum_{j=1}^m \beta_j = \beta, \quad 0 < \beta \leq 1, \quad j = 1, \cdots, m \quad (6)$$

则称不等式 $A \leq B$, $A < B$ (或 $B \leq A$, $B < A$) 为未确知事件.

定义 4. 称如下形式的规划:

$$\begin{aligned} \max \quad & \bar{f} \\ \text{s.t.} \quad & \text{Cr}\{f(\mathbf{x}, A) \geq \bar{f}\} \geq \beta \\ & \text{Cr}\{g_j(\mathbf{x}, B_j) \leq 0, \quad j = 1, \cdots, p\} \geq \alpha \end{aligned} \quad (7)$$

为未确知机会约束规划. 其中 $\mathbf{x}$ 为决策向量, A, B_j ($j = 1, \cdots, p$) 为未确知参数向量; $f(\mathbf{x}, A)$ 为决策向量 $\mathbf{x}$ 的函数, 且为未确知参数向量的有理函数, 称为目标函数; $g_j(\mathbf{x}, B_j)$ ($j = 1, \cdots, p$) 为决策向量 $\mathbf{x}$ 的函数, 且为未确知参数向量 B_j ($j = 1, \cdots, p$) 的有理函数, 称为约束函数; α, β ($\alpha, \beta \in (0, 1]$) 为事先给定的对约束条件和目标函数的置信水平; $\text{Cr}\{\cdot\}$ 为未确知事件 $\{\cdot\}$ 的可信度 (Credible degree).

2 未确知支持向量机

在以上预备知识基础上, 我们研究未确知支持向量机. 若支持向量机的训练点的输出中含有未确知信息, 则可以将未确知信息转化为未确知数:

$$A = [[y_1, y_n], f(y)] \quad (8)$$

其中

$$f(y) = \begin{cases} \alpha_j, & y = y_j, \quad j = 1, \cdots, n \\ 0, & y \neq y_j, \quad y \in [y_1, y_n] \end{cases} \quad (9)$$

因此支持向量机的训练集为

$$S = \{(\mathbf{x}_1, A_1), (\mathbf{x}_2, A_2), \cdots, (\mathbf{x}_l, A_l)\} \quad (10)$$

其中, $\mathbf{x}_i \in \mathbf{R}^n$, A_i 为未确知数. $(\mathbf{x}_i, A_i)$ ($i = 1, \cdots, l$) 称为未确知训练点, S 称为未确知训练集.

对于线性未确知训练集, 我们可将未确知分类问题 (训练点中含有未确知信息的分类问题) 转化为求解如下的未确知机会约束规划问题:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \text{Cr}\{A_i((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1\} \geq \lambda, \quad i = 1, \cdots, l \end{aligned} \quad (11)$$

其中, $(\mathbf{w}, b)^T$ 为决策变量, A_i ($i = 1, \dots, l$) 为未确知训练集 (10) 中的未确知数, $\text{Cr}\{\cdot\}$ 为未确知事件 $\{\cdot\}$ 的可信度, λ ($0 < \lambda \leq 1$) 为置信水平.

下面求解未确知机会约束规划 (11), 即求出式 (11) 的确知等价规划.

定理 1. 在置信水平 λ ($0 < \lambda \leq 1$) 下, 未确知机会约束规划式 (11) 的确知等价规划为

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & E_i((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1, \quad i = 1, \dots, l \end{aligned} \quad (12)$$

其中

$$E_i = \max\{y_{ik} | F(y_{i(k-1)}) \cong \alpha - \lambda, k = 2, \dots, n\} > 0 \quad (13)$$

或

$$E_i = \min\{y_{ik} | F(y_{ik}) \cong \lambda, k = 1, \dots, n\} < 0 \quad (14)$$

$F(y_{i(k-1)})$ 为未确知数 (分布函数型) A_i 的分布函数在 $y_{i(k-1)}$ 处的值, $F(y_{ik})$ 为未确知数 (分布函数型) A_i 的分布函数在 y_{ik} 处的值.

证明. 对于式 (11) 中的未确知约束条件:

$$\text{Cr}\{A_i((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1\} \geq \lambda, \quad i = 1, \dots, l \quad (15)$$

其中, $\mathbf{w} = (w_1, \dots, w_n) \in \mathbf{R}^n$, $b \in \mathbf{R}$, $\mathbf{x}_i = (x_{i1}, \dots, x_{in}) \in \mathbf{R}^n$, $A_i = [[y_{i1}, y_{in}], f_i(y)]$,

$$f_i(y) = \begin{cases} \alpha_{ij}, & y = y_{ij}, \quad j = 1, \dots, n \\ 0, & \text{其他} \end{cases}$$

$$0 < \sum_{j=1}^n \alpha_{ij} \leq 1, \quad 0 < \alpha_{ij} \leq 1, \quad j = 1, \dots, n$$

$$\text{令 } \mathbf{w}' = (b, w_1, \dots, w_n), \quad \mathbf{x}'_i = (1, x_{i1}, \dots, x_{in}),$$

则

$$\begin{aligned} (\mathbf{w} \cdot \mathbf{x}_i) + b &= w_1 x_{i1} + \dots + w_n x_{in} + b = \\ &= (b, w_1, \dots, w_n)(1, x_{i1}, \dots, x_{in})^T = \\ &= (\mathbf{w}' \cdot \mathbf{x}'_i) \end{aligned} \quad (16)$$

故可将式 (15) 表示为

$$\text{Cr}\{A_i(\mathbf{w}' \cdot \mathbf{x}'_i) \geq 1\} \geq \lambda, \quad i = 1, \dots, l \quad (17)$$

即

$$\text{Cr}\{1 - A_i(\mathbf{w}' \cdot \mathbf{x}'_i) \leq 0\} \geq \lambda, \quad i = 1, \dots, l \quad (18)$$

这里, $(\mathbf{w}' \cdot \mathbf{x}'_i) \neq 0$ (若 $(\mathbf{w}' \cdot \mathbf{x}'_i) = 0$, 则样本点在分离超平面上, 此时对于整个分类问题不起作用, 所以

可以只考虑不在分离超平面上的样本点的情况, 即 $(\mathbf{w}' \cdot \mathbf{x}'_i) \neq 0$). 因此可得到:

$$\text{Cr}\left\{(\mathbf{w}' \cdot \mathbf{x}'_i) \left[\frac{1}{(\mathbf{w}' \cdot \mathbf{x}'_i)} - A_i\right] \leq 0\right\} \geq \lambda, \quad i = 1, \dots, l \quad (19)$$

由未确知事件可信度的定义^[5, 20], 式 (19) 可写为

$$\begin{cases} \text{Cr}\{\frac{1}{(\mathbf{w}' \cdot \mathbf{x}'_i)} - A_i \leq 0\} \geq \lambda, & \mathbf{w}' \cdot \mathbf{x}'_i > 0 \\ \text{Cr}\{A_i - \frac{1}{(\mathbf{w}' \cdot \mathbf{x}'_i)} \leq 0\} \geq \lambda, & \mathbf{w}' \cdot \mathbf{x}'_i < 0 \end{cases} \quad (20)$$

根据文献 [20] 得, 当 $(\mathbf{w}' \cdot \mathbf{x}'_i) > 0$ 时, $\text{Cr}\{\frac{1}{(\mathbf{w}' \cdot \mathbf{x}'_i)} - A_i \leq 0\} \geq \lambda$ 等价于

$$\frac{1}{(\mathbf{w}' \cdot \mathbf{x}'_i)} \leq P_i \quad (21)$$

其中, $P_i = \max\{y_{ik} | F(y_{i(k-1)}) \cong \alpha - \lambda, k = 2, \dots, n\}$. 当 $(\mathbf{w}' \cdot \mathbf{x}'_i) < 0$ 时, $\text{Cr}\{A_i - \frac{1}{(\mathbf{w}' \cdot \mathbf{x}'_i)} \leq 0\} \geq \lambda$ 等价于

$$\frac{1}{(\mathbf{w}' \cdot \mathbf{x}'_i)} \geq H_i \quad (22)$$

其中, $H_i = \min\{y_{ik} | F(y_{ik}) \cong \lambda, k = 1, \dots, n\}$. 根据不等式性质, 将式 (21) 和式 (22) 合并为

$$(\mathbf{w}' \cdot \mathbf{x}'_i) E_i \geq 1 \quad (23)$$

其中, $E_i = P_i = \max\{y_{ik} | F(y_{i(k-1)}) \cong \alpha - \lambda, k = 2, \dots, n\} > 0$ 或 $E_i = H_i = \min\{y_{ik} | F(y_{ik}) \cong \lambda, k = 1, \dots, n\} < 0$. 将式 (16) 代入式 (23), 得:

$$E_i((\mathbf{w}' \cdot \mathbf{x}'_i) + b) \geq 1 \quad (24)$$

即在置信水平 λ ($0 < \lambda \leq 1$) 下, 未确知机会约束规划式 (11) 的确知等价规划为式 (12). $\square$

定理 2. 二次规划式 (12) 的对偶规划为

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l E_i E_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{j=1}^l \alpha_j \\ \text{s.t.} \quad & \sum_{i=1}^l E_i \alpha_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, \dots, l \end{aligned} \quad (25)$$

证明. 略. $\square$

规划 (25) 为一个凸二次规划, 解得其最优解为 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$, 则令 $\mathbf{w}^* = \sum_{i=1}^l E_i \alpha_i^* \mathbf{x}'_i$, $b^* = \frac{1}{E_i} - \sum_{i=1}^l \frac{1}{E_i} \alpha_i^* (\mathbf{x}_i \cdot \mathbf{x}_j)'$, 其中 j 对应的 α_j^* 非零. 得最优化分类函数:

$$f(x) = \text{sgn}((\mathbf{w}^* \cdot \mathbf{x}) + b^*) =$$

$$\operatorname{sgn} \left(\sum_{i=1}^l E_i \alpha_i^* (\mathbf{x} \cdot \mathbf{x}_i) + b^* \right) \quad (26)$$

引入惩罚因子和核函数后, 通过以上讨论, 我们得出算法如下:

算法 1. 未确知支持向量机

1) 给定未确知训练集 (10), 选择适当的置信水平 λ ($0 < \lambda \leq 1$), 根据 $E_i = \max\{y_{ik} | F(y_{i(k-1)}) \cong \alpha - \lambda, k = 2, \dots, n\}$ 或 $E_i = \min\{y_{ik} | F(y_{ik}) \cong \lambda, k = 1, \dots, n\}$, 求出 E_i , 其中, $F(y_{i(k-1)})$ 为未确知数 (分布函数型) A_i 的分布函数在 $y_{i(k-1)}$ 处的值. $F(y_{ik})$ 为未确知数 (分布函数型) A_i 的分布函数在 y_{ik} 处的值.

2) 选取适当的核函数 $K(\mathbf{x}, \mathbf{x}')$ 和适当的惩罚参数 $C > 0$, 构造最优化问题:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l E_i E_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{j=1}^l \alpha_j \\ \text{s.t.} \quad & \sum_{i=1}^l E_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \dots, l \end{aligned} \quad (27)$$

并求出最优解 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$.

3) 选择 α^* 的一个正分量 α_j^* , 据此计算 $b^* = \frac{1}{E_j} - \sum_{i=1}^l \frac{1}{E_i} \alpha_i^* K(\mathbf{x}_i, \mathbf{x}_j)$.

4) 构造最优分类器函数

$$f(x) = \operatorname{sgn} \left(\sum_{i=1}^l E_i \alpha_i^* K(\mathbf{x}, \mathbf{x}_i) + b^* \right) \quad (28)$$

注 1. 未确知支持向量机中置信水平 λ ($0 < \lambda \leq 1$) 的确定可看作参数选择问题, 可以采用参数选择方法, 如: LOO (Leave-one-out) 误差、LOO 误差界等^[7-9].

注 2. 规划 (27) 的最优解 $(\alpha_1^*, \dots, \alpha_l^*)^T$ 中只有一部分 (通常是少部分) α_i^* 不为零, 而它们所对应的未确知训练点的输入 $\mathbf{x}_i$ 称为未确知支持向量. 所以最优分类函数可表示为

$$f(x) = \operatorname{sgn} \left(\sum_{\mathbf{x}_i \in \text{USV}} E_i \alpha_i^* K(\mathbf{x}, \mathbf{x}_i) + b^* \right) \quad (29)$$

其中, USV 为未确知支持向量组成的集合.

注 3. 若未确知训练集中所有未确知训练点输出全为实数 1 或 -1, 则未确知训练集退化为标准训练集. 此时未确知支持向量机变为标准两类分类问题的支持向量机. 否则, 训练集中含有未确知信息, 未确知支持向量机可理解为半监督学习方法.

注 4. 选择不同的核函数, 可以生成不同的未确知支持向量机, 常用的有以下几种:

- 1) 线性未确知支持向量机: $K(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$;
- 2) 多项式未确知支持向量机: $K(\mathbf{x}, \mathbf{x}') = [(\mathbf{x} \cdot \mathbf{x}') + c]^d, c \geq 0$;
- 3) 径向基函数未确知支持向量机: $K(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2)$;
- 4) 二层神经网络未确知支持向量机: $K(\mathbf{x}, \mathbf{x}') = \tanh[k(\mathbf{x} \cdot \mathbf{x}') - c], k > 0$.

3 实验结果

3.1 UCI 数据

首先, 用 6 个 UCI (具体见 http://www.ics.uci.edu/mllearn/MLR_epository.html) 公共数据集上的实验验证本文提出的算法的有效性. 针对 UCI 中两类分类问题的数据集, 我们将其分为训练集和预测集两部分 (比例为 4:1). 对训练集, 随机去掉一部分 (50%) 样本的标签, 并根据 k 近邻 (这里取 10 近邻) 给出标签的未确知信息, 具体做法为: 选取无标签样本周围的 10 近邻, 其中正类样本、负类样本和无标签样本个数除以 10 即为该样本的正、负和未确知信息, 进而给出未确知数, 并以此作为未确知支持向量机的训练集; 同时, 将对比半监督支持向量机算法^[23-24]和模糊支持向量机^[10]在该训练集和预测集上的实验结果. 本文中惩罚参数 C 及核参数 σ (径向基核函数) 的选取范围为 $[10^{-5}, 10^6]$, $\lambda \in (0, 1]$. 在本文所有实验中, 我们使用 10 折交叉验证选择最优参数, 取得最优参数后对训练集训练并对预测集进行预测.

表 1 分别给出了半监督支持向量机、模糊支持向量机和未确知支持向量机采用线性和径向基核函数的准确率. 从表 1 中可以得出, 在采用以上线性函数的情况下, 未确知支持向量机的准确率均明显高于模糊支持向量机和半监督支持向量机, 从而说明与半监督支持向量机和模糊支持向量机相比, 未确知支持向量机更有效地利用了分类问题中的不确定性信息, 从而具有更好的分类性能. 非线性的结果与线性相似, 从而更进一步说明了未确知支持向量机的有效性.

3.2 应用实例

下面以人体亚健康诊断中是否处于中度以上亚健康状态的应用实例^[25-26]来进一步验证未确知支持向量机 (算法) 的有效性.

这里选用健康人和处于中度以上亚健康状态的人作为研究对象, 以被试者脉搏信号和亚健康自测表测得脉象样本的原始分类作为实验数据^[26]. 以无躯体疾病和无精神障碍的人为测试对象, 使用集成化数字脉搏传感器, 检测被测试者左关部位的脉搏信号, 并用数字低通滤波器滤除高频干扰, 然后选取

表 1 UCI 数据集分类器准确率对比试验结果表 (%)

Table 1 Comparison of testing percentage accuracy of classifiers on UCI data set (%)

数据集	半监督支持向量机		模糊支持向量机		未确知支持向量机	
	线性	非线性	线性	非线性	线性	非线性
Sonar (208 × 60)	73.81	83.33	79.50	84.47	80.95	90.48
WPBC (198 × 34)	77.50	80.00	82.80	83.05	82.50	85.00
Hepatitis (155 × 19)	67.74	70.97	72.05	77.25	80.65	80.65
Australian (690 × 14)	72.46	76.09	70.25	73.69	79.71	80.44
Pima-Indian (768 × 8)	70.13	70.78	67.48	78.26	74.68	77.92
Votes (435 × 16)	91.95	93.10	94.21	95.25	93.10	95.40

一个完整的脉搏波, 采用 Welch 法^[26] 进行功率谱估计, 最后使用多导睡眠图监测 (Polysomnogram, PSG) 提取 0 Hz ~ 30 Hz 频段进行分析, 选取功率谱重心、重心频率功率、谱峰值和峰值频率作为特征量. 以测试对象在亚健康自测表上测得的结果为类别标签. 但是由于亚健康自测表含有大量决策者的主观因素, 含有一些未确知信息. 对于这些未确知信息, 如果把它们笼统地视为确知信息处理的话, 可能会造成很大的误差, 甚至错误. 因此对上述数据, 我们用未确知支持向量机对其进行分类.

实验中, 我们选取 60 个样本, 健康组 30 例, 亚健康组 30 例, 并将这 60 个样本分为两组, 一组作为训练, 包括 18 个健康样本点和 17 个亚健康样本点, 其中这 35 个训练样本点的类别标签为未确知数. 其中表 2 给出了 10 个含有未确信息的样本点, 对于含有未确知信息的类标签, 用未确知数来表示. 进一步, 根据医学专家提供的未确知信息, 由定义 1 构造 10 个未确知数的可信度分布密度函数, 见表 2 和表 3. 其中一组作为测试, 包括 12 个亚健康样本点和 13 个健康样本点.

因此, 由表 2 和表 3, 可将含有未确知信息的训练集表示为如下形式: $S = \{(\mathbf{x}_1, 1), (\mathbf{x}_2, 1), \dots, (\mathbf{x}_{25}, -1), (\mathbf{x}_{26}, A_1), (\mathbf{x}_{27}, A_2), \dots, (\mathbf{x}_{35}, A_{10})\}$.

对以上的数据集, 首先, 对数据集进行归一化. 然后, 代入算法搜索并得到最优的惩罚参数 $C = 100$, $\lambda = 0.7$ 和 $\sigma = 2.85$. 并可求得对应于最优参数下的未确知支持向量如下:

$$\begin{aligned}\mathbf{x}_9 &= (61\ 135, 1.2048, 53.481, 0.75) \\ \mathbf{x}_{12} &= (8\ 582.9, 1.5852, 44.021, 1) \\ \mathbf{x}_{15} &= (37\ 190, 1.1363, 52.595, 0.625) \\ \mathbf{x}_{19} &= (30\ 509, 1.2106, 50.482, 0.75) \\ \mathbf{x}_{25} &= (22\ 904, 1.2604, 47.408, 0.875)\end{aligned}$$

表 2 含有未确知信息的亚健康诊断数据

Table 2 The subhealth diagnostic data set with unascertained information

类型	功率谱	重心	功率谱	峰值
	重心值	频率	峰值	频率
A1	80 585	1.2555	45.561	0.75
A2	81 433	1.331	52.32	0.875
A3	32 439	1.1092	51.583	0.75
A4	128 660	1.2245	53.626	0.875
A5	32 439	1.1092	51.583	0.75
A6	62 253	1.3693	50.822	0.875
A7	26 358	1.1702	46.461	0.75
A8	8 209.9	1.1402	42.708	0.75
A9	128 660	1.2245	53.626	0.875
A10	32 439	1.0092	51.583	0.75

注: A_i 表示未确知信息.

为了显示未确知支持向量机 (算法) 处理未确知分类问题的优越性, 将半监督支持向量机和模糊支持向量机用同样的数据进行训练和测试, 对比结果如表 4 所示. 表 4 中, 三种 SVM 模型使用的参数均为各自的最优参数. 由表 4 可以看出, 采用半监督支持向量机和模糊支持向量机进行训练的时间分别为 356 s 和 42.55 s, 测试的准确率分别为 88.00 % 和 90.00 %; 采用未确知支持向量机进行训练的时间为 35.71 s, 测试的准确率为 96.00 %. 半监督支持向量机和模糊支持向量机不仅在测试准确率上逊色于未确知支持向量机, 同时训练的时间也大于未确知支持向量机的训练时间. 这是因为样本信息中含有未确知信息, 半监督支持向量机会去掉含有未确知信息的类别标签, 把它当作无标签数据, 然后预测这

些无标签数据的类别标签, 这样往往会产生误差, 使得训练精度大打折扣. 同时这样做的复杂度和训练时间也会大于未确知支持向量机.

表 3 含有未确知信息对应的未确知数
Table 3 The value corresponding to
the unascertained information

$f_1(y) = \begin{cases} 0.2, & y = -1 \\ 0.7, & y = 1 \\ 0, & \text{其他} \end{cases}$	$f_2(y) = \begin{cases} 0.4, & y = -1 \\ 0.6, & y = 1 \\ 0, & \text{其他} \end{cases}$
$f_3(y) = \begin{cases} 0.7, & y = -1 \\ 0.3, & y = 1 \\ 0, & \text{其他} \end{cases}$	$f_4(y) = \begin{cases} 0.2, & y = -1 \\ 0.8, & y = 1 \\ 0, & \text{其他} \end{cases}$
$f_5(y) = \begin{cases} 0.9, & y = -1 \\ 0.1, & y = 1 \\ 0, & \text{其他} \end{cases}$	$f_6(y) = \begin{cases} 0.2, & y = -1 \\ 0.7, & y = 1 \\ 0, & \text{其他} \end{cases}$
$f_7(y) = \begin{cases} 0.8, & y = -1 \\ 0.2, & y = 1 \\ 0, & \text{其他} \end{cases}$	$f_8(y) = \begin{cases} 0.7, & y = -1 \\ 0.3, & y = 1 \\ 0, & \text{其他} \end{cases}$
$f_9(y) = \begin{cases} 0.1, & y = -1 \\ 0.9, & y = 1 \\ 0, & \text{其他} \end{cases}$	$f_{10}(y) = \begin{cases} 0.6, & y = -1 \\ 0.3, & y = 1 \\ 0, & \text{其他} \end{cases}$

表 4 实验结果
Table 4 Experimental results

分类器	参数	准确率 (%)	训练 时间 (s)
未确知 SVM	$\sigma = 2.85, \lambda = 0.7$ $C = 100$	96.00	35.71
半监督 SVM	$\sigma = 8, C = 50$	88.00	356
模糊 SVM	$\sigma = 1, C = 0.1$	90.00	42.55

通过以上实验可以看出, 在解决样本信息中含有未确知信息的学习问题时, 未确知支持向量机无论是从学习的质量还是学习的速度来说, 都要比模糊支持向量机和半监督支持向量机优越. 特别是未确知支持向量机在置信水平 λ 的控制和调节下, 更能提高分类器性能.

4 结论

针对未确知分类问题, 本文在标准支持向量机(算法)基础上, 以未确知数学和未确知规划为工具, 根据未确知信息的特点, 建立了未确知支持向量机(算法). 并且研究了未确知支持向量机(算法)在公

共数据集上的有效性以及在人体亚健康诊断中的应用, 对比实验表明本文提出的算法效果良好.

References

- 1 Wang Guang-Yuan. Unascertained information and its mathematical treatment. *Journal of Harbin University of Civil Engineering and Architecture*, 1990, **23**(4): 1–9
(王光远. 未确知信息及其数学处理. 哈尔滨建筑工程学院学报, 1990, **23**(4): 1–9)
- 2 Liu Kai-Di, Wu He-Qin, Pang Yan-Jun. *Mathematical Treatment and Application of Uncertain Information*. Beijing: Science Press, 1999
(刘开第, 吴和琴, 庞彦军. 不确定性信息数学处理及应用. 北京: 科学出版社, 1999)
- 3 Yang Z M. Review of the monograph mathematical treatment and application of uncertain information. *Chinese Science Bulletin*, 2001, **46**(7): 615–616
- 4 Liu Kai-Di, Cao Qing-Kui, Pang Yan-Jun. A method of fault diagnosis based on unascertained set. *Acta Automatica Sinica*, 2004, **30**(5): 747–756
(刘开第, 曹庆奎, 庞彦军. 基于未确知集合的故障诊断方法. 自动化学报, 2004, **30**(5): 747–756)
- 5 Wu Qing-E. *The Treatment Theory, Method and Application of Uncertain Information*. Beijing: Science Press, 2009
(吴青娥. 不确定信息处理理论、方法及其应用. 北京: 科学出版社, 2009)
- 6 Li Ming, Wang Yan, Nian Fu-Zhong. *Intelligent Information Processing and Application*. Beijing: Electronic Industry Press, 2010
(李明, 王燕, 年福忠. 智能信息处理与应用. 北京: 电子工业出版社, 2010)
- 7 Vapnik V. N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995
- 8 Deng Nai-Yang, Tian Ying-Jie. *Theory, Algorithm and Extension of Support Vector Machine*. Beijing: Science Press, 2009
(邓乃扬, 田英杰. 支持向量机——理论、算法与拓展. 北京: 科学出版社, 2009)
- 9 Zhang Xue-Gong. Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 2000, **26**(1): 32–42
(张学工. 关于统计学习理论与支持向量机. 自动化学报, 2000, **26**(1): 32–42)
- 10 Lin C F, Wang S D. Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, 2002, **13**(2): 464–471
- 11 Tao Q, Wang J. A new fuzzy support vector machine based on the weighted margin. *Neural Procession Letters*, 2004, **20**(3): 139–150
- 12 Ji A B, Pang J H, Qiu H J. Support vector machine for classification based on fuzzy training data. *Expert Systems with Applications*, 2010, **37**(4): 3495–3498
- 13 Yang Zhi-Min, Deng Nai-Yang. Fuzzy support vector classification based on possibility theory. *Pattern Recognition and Artificial Intelligence*, 2007, **20**(1): 7–14
(杨志民, 邓乃扬. 基于可能性理论的模糊支持向量分类机. 模式识别与人工智能, 2007, **20**(1): 7–14)
- 14 Yang Z M, Yang X, Liu G L. A fuzzy optimization method for data mining. *Journal of Convergence Information Technology*, 2010, **5**(5): 86–91

- 15 Zadeh L A. Fuzzy sets. *Information and Control*, 1965, **8**(3): 338–353
- 16 Zadeh L A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1978, **1**(1): 3–28
- 17 Peng Zu-Zeng, Sun Wen-Yu. *Fuzzy Mathematics and Its Applications*. Wuhan: Wuhan University Press, 2007
(彭祖赠, 孙温玉. 模糊数学及其应用. 武汉: 武汉大学出版社, 2007)
- 18 Xie Ji-Jian, Liu Cheng-Ping. *The Method and Applications of Fuzzy Mathematics*. Wuhan: Huazhong University of Science and Technology Press, 2006
(谢季坚, 刘承平. 模糊数学方法及其应用. 武汉: 华中科技大学出版社, 2006)
- 19 Yang Zhi-Min, Deng Nai-Yang. Unascertained chance constrained programming. *Systems Engineering*, 2004, **22**(4): 11–14
(杨志民, 邓乃扬. 未确知机会约束规划. 系统工程, 2004, **22**(4): 11–14)
- 20 Yang Zhi-Min, Liu Guang-Li. *The Method and Applications of Uncertain Support Vector Machine*. Beijing: Science Press, 2012
(杨志民, 刘广利. 不确定性支持向量机 - 算法及应用. 北京: 科学出版社, 2012)
- 21 Shivaswamy P K, Bhattacharyya C, Smola A J. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 2006, **7**: 1283–1314
- 22 Li Xiao-Qing, Zhou Chang-Yin, Wang Yan-Chao. The method of unascertained chance constrained programming under the conditions of linear interdependence. *Journal of Tangshan College*, 2010, **23**(6): 10–12
(李小青, 周长银, 王延朝. 线性相依条件下未确知机会约束规划的解法. 唐山师范学院学报, 2010, **23**(6): 10–12)
- 23 Yang Xu-Bing, Pan Zhi-Song, Chen Song-Can. Semi-supervised proximal support vector machine via generalized eigenvalues. *Pattern Recognition and Artificial Intelligence*, 2009, **22**(3): 349–353
(杨绪兵, 潘志松, 陈松灿. 半监督型广义特征值最接近支持向量机. 模式识别与人工智能, 2009, **22**(3): 349–353)
- 24 Bennett K P, Demiriz A. Semi-supervised support vector machines. In: *Advances in Neural Information Processing Systems*. 1999, **11**: 368–374
- 25 Yang Feng-Xia. Recognition of sub-health based on support vector machine. *Modern Electronics Technique*, 2009,

30(20): 167–170

(杨凤霞. 基于支持向量机的亚健康状态识别. 现代电子技术, 2009, **30**(20): 167–170)

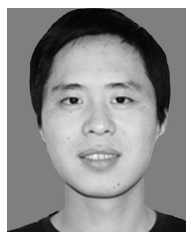
- 26 Yuan Hui. The Characteristics Information Analysis of the Pulse Wave and the Study of Arteriosclerosis Diagnostic [Master dissertation], Shandong University, China, 2005
(元慧. 脉搏波的特征信息分析和动脉硬化诊断研究 [硕士学位论文], 山东大学, 中国, 2005)



杨志民 教授. 2005 年获得中国农业大学运筹与优化专业博士学位. 主要研究方向为不确定性信息数学处理, 数据挖掘与支持向量机研究.

E-mail: yzm9966@126.com

(**YANG Zhi-Min** Professor. He received his Ph.D. degree from China Agriculture University in 2005. His research interest covers uncertainty mathematics, data mining, and support vector machines.)



邵元海 讲师. 2011 年获得中国农业大学运筹与优化专业博士学位. 主要研究方向为数据挖掘, 机器学习与支持向量机研究. 本文通信作者.

E-mail: shaoyuanhai21@163.com

(**SHAO Yuan-Hai** Lecturer. He received his Ph.D. degree from China Agriculture University in 2011. His research interest covers data mining, machine learning, and support vector machines. Corresponding author of this paper.)



梁静 助理经济师. 2011 年获得浙江工业大学应用数学专业硕士学位. 主要研究方向为不确定性信息数学处理和支持向量机.

E-mail: liangjing_0914@yahoo.com.cn

(**LIANG Jing** Assistant economist. She received her bachelor degree from Zhejiang University of Technology in 2011. Her research interest covers uncertainty mathematics and support vector machines.)