

面向分类去噪问题的模糊支持向量机新算法

张瑞^{1,2}, 马逸尘¹, 段现报¹

(1. 西安交通大学理学院, 710049, 西安; 2. 山东理工大学数学与信息学院, 255049, 淄博)

摘要: 针对面向分类去噪问题, 提出了一种新的模糊支持向量机算法(ν -FSVM), 并给出了通过无穷次连续可微函数建立模糊关系的方法. 该方法能对训练集中的点自动赋予模糊关系, 并且对带有噪声的点和孤立的点赋予较小的模糊关系. 与传统的 ν 支持向量机比较, 该算法通过建立训练集的模糊关系, 能够大大减小噪声对分类的影响, 从而提高分类精度, 减少误差.

关键词: 支持向量机; 分类; ν 支持向量机

中图分类号: TP181 **文献标识码:** A **文章编号:** 0253-987X(2007)12-1414-04

New Algorithm of Fuzzy Support Vector Machine for Classification with Outliers

Zhang Rui^{1,2}, Ma Yichen¹, Duan Xianbao¹

(1. School of Sciences, Xi'an Jiaotong University, Xi'an 710049, China; 2. School of Mathematics and Information Science, Shandong University of Technology, Zibo 255049, China)

Abstract: A fuzzy support vector machine learning approach, namely ν -FSVM is presented to eliminate classification noise. And the fuzzy relations among the training data points are also created with an infinite continue differentiable function, which enables to automatically assign the fuzzy relationships to the training data points, and assign small fuzzy relationships to the points with noises or outliers. Comparing with the traditional ν -support vector machine, the experimental results on benchmark datasets show that the presented approach reduces the outlier effects to improve the classification performance.

Keywords: support vector machine; classification; ν -support vector machine

支持向量机(SVM)是一种新的通用的机器学习方法, 它的理论基础是 Vapnik 等提出的统计学习理论, 其基本思想是通过非线性变换将输入空间变换到一个高维空间, 并在这个新空间中求取最优线性分类面. 由于最优分类面仅依赖于训练集中少量的点(称之为支持向量), 对噪声非常敏感^[1], 解决这个问题的一种方法是对训练集进行预处理, 去除噪声^[2], 但由于缺乏噪声的足够信息, 这种方法实际操作难度很大.

为了解决这个问题, 文献[3-5]提出了一种称之为模糊支持向量机(FSVM)的方法, 该方法通过对训练集中的点建立模糊关系, 其目的在于根据每个

数据点对分类结果的不同影响, 为其赋予不同的值, 以降低噪声对分类结果的影响.

本文基于分类去噪问题提出了一种模糊的 ν 支持向量机算法(ν -FSVM), 并给出了通过无穷次连续可微函数建立模糊关系的方法. 实验结果表明, 该方法能够大大减少噪声对分类的影响, 从而提高分类精度, 减少误差.

1 模糊的 ν 支持向量机新算法

1.1 ν 支持向量机(ν -SVM)

假设给定训练集

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_l, y_l)\}$$

其中 $\mathbf{x}_i \in R^N, y_i \in \{-1, 1\}, i=1, 2, \dots, l$. ν -SVM 的基本思想是:通过非线性变换 ϕ 将 N 维输入的样本点 $\{\mathbf{x}_i\}$ 从原空间 R^N 影射到高维特征空间 H , 用 $\langle \cdot, \cdot \rangle$ 表示内积, 得到新的集合 $\{(\phi(\mathbf{x}_1), y_1), \dots, (\phi(\mathbf{x}_l), y_l)\}$, 由此构造最优线性超平面. 决策函数为

$$f(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b)$$

ν -SVM 需要求解下面的优化问题

$$\left. \begin{aligned} \min \quad & \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle - \nu \rho + \frac{1}{l} \sum_{i=1}^l \xi_i \\ \text{s. t.} \quad & y_i (\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b) \geq \rho - \xi_i \\ & \xi_i \geq 0, 1 \leq i \leq l, \rho \geq 0 \end{aligned} \right\} \quad (1)$$

式中: $0 \leq \nu \leq 1$.

1.2 ν -FSVM

在传统的 SVM 理论中, 训练过程对于那些远离它们所属类的训练点是十分敏感的, 但在 SVM 的训练过程中, 所有的训练点都被平等对待. 这样就导致了 SVM 对某些特殊情形的点过分敏感, 例如孤立点与噪声, 这种情形即是所谓的过学习现象.

在很多实际应用问题中, 不同训练点对分类结果的影响是不同的. 一般来说, 训练集中存在某些点对分类结果的影响很大, 同时也存在一些点对分类结果的影响很小, 甚至是微不足道的. 因此, 在处理分类问题时, 必须将那些重要的点正确分类, 并且可以忽略那些微不足道的点, 比如说带有噪声的点和距离类中心很远的孤立点(该两类点此处统称为噪声).

假设给出了一个带有类别标号和模糊关系的训练集 $S = \{(\mathbf{x}_1, y_1, s_1), \dots, (\mathbf{x}_l, y_l, s_l)\}$, 每一个训练样本 $\mathbf{x}_i \in R^N$ 都给出了与其对应的类别标号 $y_i \in \{-1, 1\}$, 以及模糊关系 $\sigma \leq s_i \leq 1$, 其中 $\sigma > 0$ 是一个充分小的正数. 因为模糊关系 s_i 对应样本点 $\mathbf{x}_i$ 属于某一类的重要程度, 在 ν -SVM 中参数 ξ_i 可以看成是 $\mathbf{x}_i$ 对应的误差, 那么 $s_i \xi_i$ 就是对应项带有权重 s_i 的误差. 最优分类超平面可由下面二次优化问题的解给出

$$\left. \begin{aligned} \min \quad & \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle - \nu \rho + \frac{1}{l} \sum_{i=1}^l s_i \xi_i \\ \text{s. t.} \quad & y_i (\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b) \geq \rho - \xi_i \\ & \xi_i \geq 0, 1 \leq i \leq l, \rho \geq 0 \end{aligned} \right\} \quad (2)$$

可以根据式(2), 构造 Lagrangian 函数

$$L(\mathbf{w}, b, \xi, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \nu \rho +$$

$$\frac{1}{l} \sum_{i=1}^l s_i \xi_i - \sum_{i=1}^l \beta_i \xi_i -$$

$$S \rho - \sum_{i=1}^l \alpha_i [y_i (\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b) - \rho + \xi_i] \quad (3)$$

式中: $\alpha_i \geq 0; \beta_i \geq 0; i=1, \dots, l; S \geq 0$.

分别求关于 $\mathbf{w}, b, \xi_i, \rho$ 的偏导数, 在鞍点处应满足

$$\frac{\partial L(\mathbf{w}, b, \xi, \alpha)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^l \alpha_i y_i \phi(\mathbf{x}_i) = 0 \quad (4)$$

$$\frac{\partial L(\mathbf{w}, b, \xi, \alpha)}{\partial \xi_i} = \frac{1}{l} s_i - \alpha_i - \beta_i = 0 \quad (5)$$

$$i = 1, \dots, l$$

$$\frac{\partial L(\mathbf{w}, b, \xi, \alpha)}{\partial b} = - \sum_{i=1}^l \alpha_i y_i = 0 \quad (6)$$

$$\frac{\partial L(\mathbf{w}, b, \xi, \alpha)}{\partial \rho} = -\nu + \sum_{i=1}^l \alpha_i - S = 0 \quad (7)$$

将式(4)~式(7)代入式(3), 则得到问题(2)的对偶形式

$$\min_{\alpha} W(\alpha) = \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (8)$$

$$\left. \begin{aligned} \text{s. t.} \quad & \sum_{i=1}^l \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq \frac{1}{l} s_i \quad i = 1, \dots, l \\ & \sum_{i=1}^l \alpha_i \geq \nu \end{aligned} \right\} \quad (9)$$

其中 $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ 称之为核函数. Karush-Kuhn-Tucker(KKT)的条件为

$$\alpha_i^* [y_i (\langle \mathbf{w}^*, \phi(\mathbf{x}_i) \rangle + b^*) - \rho + \xi_i^*] = 0 \quad (10)$$

$$1 \leq i \leq l$$

$$\xi_i^* \left(\alpha_i^* - \frac{1}{l} s_i \right) = 0 \quad 1 \leq i \leq l$$

注意到只有当 $\alpha_i = \frac{1}{l} s_i$ 时, 松弛变量 ξ_i^* 才不为 0. 要求阈值 b^* , 只要选择两个属于不同类的训练点 $\mathbf{x}_i$ 和 $\mathbf{x}_j$, 不妨设对应的类别标号分别为 $y_i = -1, y_j = 1$, 使其满足 $0 < \alpha_i^*, \alpha_j^* < \frac{1}{l} s_i$, 那么由 KKT 条件可得

$$y_i (\langle \mathbf{w}^*, \phi(\mathbf{x}_i) \rangle + b^*) - \rho = 0 =$$

$$y_j (\langle \mathbf{w}^*, \phi(\mathbf{x}_j) \rangle + b^*) - \rho$$

亦即

$$- (\langle \mathbf{w}^*, \phi(\mathbf{x}_i) \rangle + b^*) - \rho =$$

$$\langle \mathbf{w}^*, \phi(\mathbf{x}_j) \rangle + b^* - \rho$$

于是得

$$b^* = -0.5 (\langle \mathbf{w}^*, \phi(\mathbf{x}_i) \rangle + \langle \mathbf{w}^*, \phi(\mathbf{x}_j) \rangle)$$

其中

$$w^* = \sum \alpha_i^* y_i \phi(x_i)$$

综上可以得到下面的 ν -FSVM 算法.

(1) 给定训练集 $S = \{(x_1, y_1), \dots, (x_l, y_l)\}$, 其中 $x_i \in R^N, y_i \in \{-1, 1\}, i = 1, \dots, l$.

(2) 选择合适的核函数 $K(x, x')$ 和参数 ν , 求解下面的优化问题

$$\min_{\alpha} W(\alpha) = \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (11)$$

$$\text{s. t. } \left. \begin{aligned} \sum_{i=1}^l \alpha_i y_i &= 0 \\ 0 \leq \alpha_i &\leq \frac{1}{l} s_i \quad i = 1, \dots, l \\ \sum_{i=1}^l \alpha_i &\geq \nu \end{aligned} \right\} \quad (12)$$

得到解 $\alpha^* = (\alpha_1^*, \dots, \alpha_l^*)^T$.

(3) 如果 $S^+ = \{i | \alpha_i^* \in (0, \frac{1}{l}), y_i = 1\} \neq \emptyset$, 并且 $S^- = \{i | \alpha_i^* \in (0, \frac{1}{l}), y_i = -1\} \neq \emptyset$, 那么取 $i \in S^+, j \in S^-$, 计算阈值

$$b^* = -0.5 \sum_{k=1}^l \alpha_k y_k (K(x_k, x_i) + K(x_k, x_j))$$

(4) 所求的决策函数是

$$f(x) = \text{sgn}\left(\sum_{i=1}^l \alpha_i^* y_i K(x, x_i) + b^*\right)$$

2 建立训练集的模糊关系

训练集模糊关系建立的好坏,直接影响到去噪的效果.目前,有多种确定模糊关系的方法^[6-9],例如线性函数、二次函数、启发式、噪声分布等方法.本文介绍一种简单有效的去噪方法,利用无穷次连续可微函数建立训练集的模糊关系,具体方法如下

$$\text{记 } S^+ = \{x_i | y_i = 1\}; S^- = \{x_i | y_i = -1\}$$

设

$$s(z) = \begin{cases} b\sigma + b \exp\left(\frac{-1}{r^2 - \|z - c\|^2}\right) & \|z - c\| < r \\ b\sigma & \text{其他} \end{cases} \quad (13)$$

这里 $b = \left(\sigma + \exp\left(\frac{-1}{r^2}\right)\right)^{-1}$, σ 是一个充分小的正数,该函数是一个无穷次连续可微函数, $\|z - c\|$ 表示两点间的距离,图 1 和图 2 分别对应其一维和二维图形.

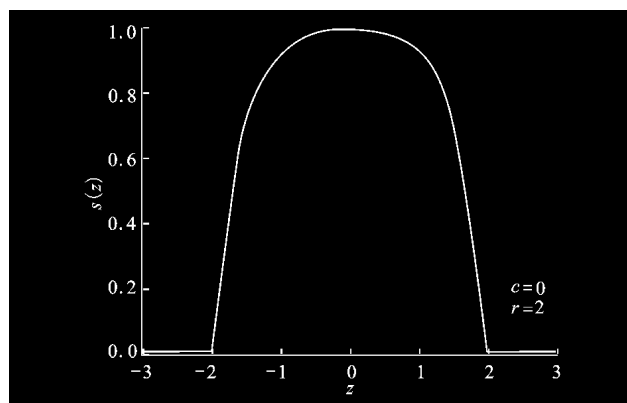


图1 一维函数 $s(z)$ 的图形

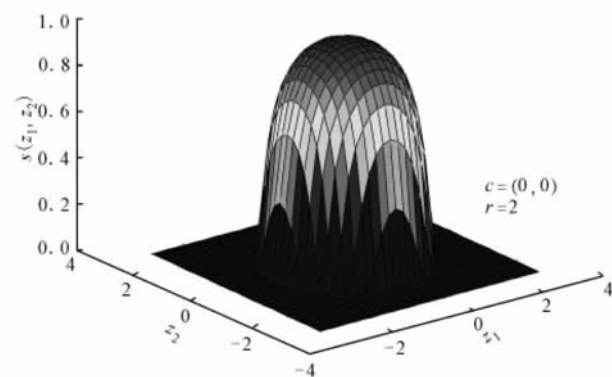


图2 二维函数 $s(z_1, z_2)$ 的图形

从图形可以看出,离中心点越近,对应的值越大,在 $\|z - c\| = r$ 上的点对应的值最小.一般情况下,噪声点离该类点的中心都比较远,或在 $\|z - c\| = r$ 上.如果 c 是某一类(正类或负类)的中心, r 是包含该类点集的最小超球面的半径,那么对噪声点按照式(13)计算得到的值(模糊关系)比较小,由此对正确分类造成的影响就会比较小.参考文献[10]求包含训练集的最小超球面 (c, r) .

3 试验和结论

下面的试验中,我们选择了多项式核作为核函数,即 $K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^2$. 在标准数据库 thyroid_train_data10, thyroid_test_data10, heart_train_data1 上对于 ν -FSVM 和传统的 ν -SVM 进行了比较,试验误差采用留一法(LOO)计算,结果见表1~表3. 模糊关系按式(13)计算,其中 c 和 r 是指包含相应类(正类或负类)的最小超球面的球心和半径, $\sigma = 0.001$. 为了便于比较,我们取 ν 分别为 0.001、0.005、0.01、0.05.

试验结果表明,该文提出的方法 ν -FSVM 与传统的 ν -SVM 相比,能够大大减少噪声对分类的影响,从而提高分类精度,减少误差.

表1 训练集为 thyroid_train_data10 的 LOO 误差

ν	LOO 误差	
	ν -SVM	ν -FSVM
0.001	0.157 1	0.078 6
0.005	0.185 7	0.114 3
0.010	0.157 1	0.107 1
0.050	0.221 4	0.064 3

表2 训练集为 thyroid_test_data10 的 LOO 误差

ν	LOO 误差	
	ν -SVM	ν -FSVM
0.001	0.266 7	0.026 7
0.005	0.280 0	0.026 7
0.010	0.280 0	0.040 0
0.050	0.266 7	0.026 7

表3 训练集为 heart_train_data1 的 LOO 误差

ν	LOO 误差	
	ν -SVM	ν -FSVM
0.001	0.359 2	0.252 9
0.005	0.364 7	0.276 5
0.010	0.364 7	0.264 7
0.050	0.382 4	0.200 0

参考文献:

- [1] Boser B E, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers [C] // Proc Fifth Ann Workshop Computational Learning Theory. New

York: ACM Press, 1992: 144-152.

- [2] Cao L J, Lee H P, Chong W K. Modified support vector novelty detector using training data with outliers [J]. Pattern Recognition Letters, 2003, 24(14): 2479-2487.
- [3] Inoue T, Abe S. Fuzzy support vector machines for pattern classification [C] // Proceedings of the International Joint Conference on Neural Networks. Piscataway: IEEE, 2001: 1449-1454.
- [4] Lin C F, Wang S D. Fuzzy support vector machines [J]. IEEE Trans Neural Networks, 2002, 13 (2): 464-471.
- [5] Lin C F, Wang S D. Training algorithms for fuzzy support vector machines with noisy data [J]. Pattern Recognition Letters, 2004, 25(14): 1647-1656.
- [6] Chen J H, Chen C S. Fuzzy kernel perceptron [J]. IEEE Trans Neural Networks, 2002, 13(6): 1364-1373.
- [7] Mill J, Inoue A. An application of fuzzy support vectors [C] // Proceedings of the 22nd North American Fuzzy Information Processing Society. Piscataway: IEEE, 2003: 302-306.
- [8] Keller J M, Hunt D J. Incorporating fuzzy membership functions into the perception algorithm [J]. IEEE Trans Patt Anal Mach Intell, 1985, 7(6): 693-699.
- [9] Tsujinishi D, Abe S. Fuzzy least squares support vector machines [C] // Proceedings of the International Joint Conference on Neural Networks. Piscataway: IEEE, 2003: 1599-1604.
- [10] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis [M]. Cambridge, UK: Cambridge University Press, 2004.

(编辑 杜秀杰)

徐宗本教授被聘为 2007 年度“973 计划”项目首席科学家

2007 年国家重点基础研究发展计划项目实施会于 9 月 26 日在京召开。我校申报的“基于视觉认知的非结构化信息处理理论与关键技术”项目获得 2007 年度“973 计划”资助,徐宗本教授被聘为该项目的首席科学家,该项目未来 5 年共可获得 3 200 万元国拨专项经费资助。

该项目将围绕解决移动主体视觉导引和复杂背景下的目标识别与跟踪两个国家重大需求共性基础问题,从认知科学、数学与信息技术交叉融合的角度,重点解决非结构化信息处理的视觉认知机理与计算模型、非结构化信息的结构化表征与计算、基于非结构化信息处理的环境感知与目标识别等关键科学问题。