计算机科学2004Vol. 31№. 1

基于模糊支持向量机的数据域描述

魏立力　龙卫江　张文修

（西安交通大学理学院信息与系统科学研究所　西安710049）

摘　要　受模糊支持向量机的启发，本文系统论述了带有模糊隶属度的数据域描述方法，称为模糊支持向量域描述。适用于数据集中的数据不完全肯定来自于假设总体的情形，不同的数据对数据集的域描述可以有不同的贡献。

关键词　支持向量机，模糊隶属度，数据域描述

Data Domain Description Based on Fuzzy Support Vector Machines

WEI Li-Li　LONG Wei-Jiang　ZHANG Wen-Xiu

(Institute of Information Sciecnce and System Science, Faculty of Science, Xi'an Jiaotong University, Xi'an 710049)

Abstract　In this paper, we reformulate the use of a data domain description method, inspired by the fuzzy support vector machine by Lin, called the fuzzy support vector domain description. This data description is suitable for applicatios in which each input point may not be fully assigned to one class. In this method, different input data can make different contributions to the domain description.

Keywords　Support vector machines, Fuzzy membership, Data domain description

1. 引言

统计学习理论[1]表明，在机器学习问题中，学习机器的期望风险可以由两部分控制：一是经验风险（训练误差），另一部分称为置信界，是学习机器 VC 维的单调增加函数。因此，机器学习过程应该在使经验风险最小的同时，还要使 VC 维尽可能小，才能保证（以一定的概率）取得小的实际风险。

以二分类问题为例，在 d 维欧氏空间中，对满足约束$\|w\| \leqslant \Delta$的超平面所构成的指示函数集$\{f(x,w,b)=\mathrm{sign}(w\cdot x+b)\}$的 VC 维 h 满足下面的界[2]：

$$h \leqslant \min([\Delta^2 R^2], d)+1, \tag{1}$$

其中 R 是包含训练样本$\{x_1, x_2, \cdots, x_l\}$的超球面的半径。

由(1)式不难发现，使 Δ^2和 R^2极小，就能够极小化 VC 维 h 的上界，从而达到控制学习机的复杂性，保证所得决策函数的推广性。其中使得 Δ^2极小，就是最大间隔分类支持向量机[3]的核心思想。而使 R^2最小，就是本文所论及的所谓数据域描述问题[4,5]。

数据集的域描述是模型式识别中继分类问题和回归问题之后的第三类问题。其基本任务是对训练数据所在的类进行描述，并且拒绝来自其它所有可能类的数据，数据域描述可用于异常值的识别和新模式的发现。

Lin[6]以区别每个样本数据对分类决策有不同的贡献为目的，引入模糊支持向量机。受 Lin 工作的启发，本文将模糊隶属度引入到数据的域描述方法[1]中，从公式推导和数值实验两个方面研究了本方法的优越性。

2. 基于模糊支持向量机的数据域描述

设有 d 维欧氏空间的训练样本$\{x_1, x_2, \cdots, x_l\}$，我们想要寻找一个能够包含所有（或大多数）训练样本的一个最小超球面，也就是要确定球心 $a \in R^d$ 和半径 $R \in R^1$，使得半径 R 最小。这样可以减少接受异常值的机会，即减少犯第二类错误的机会。

在理论上，人们总是假定每个训练样本以同等机会来自于它们所在的类总体。但在很多实际问题中，每个训练样本来自于假设类总体的置信程度是不同的，其中的一部分样本也许比另一部分更为重要或可信。我们要求在构造数据域描述超球面时，对于置信程度比较大的样本应该给予充分重视，而对于置信程度比较小的点（可能是异常值）应限制其作用。

借助于模糊集方法，我们给每个训练样本 x_i 赋予一个隶属度 $s_i(0<\sigma \leqslant s_i \leqslant 1)$以反映 x_i 的置信程序。这样训练样本集可以表示为

$$(x_1, s_1), (x_2, s_2), \cdots, (x_l, s_l) \tag{2}$$

我们的目的是寻找超球面的球心 a 和半径 R，并且最小化 R。由于 R 对离群样本点的敏感性（个别离群点可能使 R 得变得很大），我们的策略是，允许样本点可以在超球面以外，但加以惩罚。这样我们的问题就归结为如下带有不等式约束的优化问题：

$$\begin{aligned} &\min_{a,R,\xi} R^2 + C\sum_{i=1}^{l} s_i\xi_i, \\ &\text{s.t. } \|x_i - a\|^2 \leqslant R^2 + \xi_i, \\ &\quad \xi_i \geqslant 0, i=1,\cdots,l. \end{aligned} \tag{3}$$

其中 ξ_i 为松弛变量，$\|\cdot\|$为欧氏范数，C 为常数，其作用为平衡超球面的大小和异常值的数量，称为惩罚因子。

优化问题式(3)的 Lagrange 函数为

$$L(R,a,\xi,\alpha,\beta) = R^2 + C\sum_{i=1}^{l} s_i\xi_i - \sum_{i=1}^{l}\alpha_i(R^2+\xi_i-\|x_i-a\|^2) - \sum_{i=1}^{l}\beta_i\xi_i, \tag{4}$$

其中 α_i 和 β_i 为 Lagrange 乘子。式(4)分别对 R,a,ξ 求导并令其为零得：

$$\frac{\partial L}{\partial R} = 2R - 2R\sum_{i=1}^{l}\alpha_i = 0,$$

$$\frac{\partial L}{\partial a} = \sum_{i=1}^{l}\alpha_i(x_i - a) = 0,$$

$$\frac{\partial L}{\partial \xi_i} = s_i C - \alpha_i - \beta_i = 0, i=1,\cdots,l。$$

即

魏立力，龙卫江　博士生，研究统计学与机器学习。张文修　博导，研究模糊集，粗糙集，人工智能。

$$\sum_{i=1}^{l}\alpha_i=1, \tag{5}$$

$$a=\sum_{i=1}^{l}\alpha_i x_i, \tag{6}$$

$$\beta_i=s_iC-\alpha_i, i=1,\cdots,l。 \tag{7}$$

将(5)、(6)、(7)式带入 Lagrange 函数式(4)得到对偶优化问题的目标函数：

$$L_D(\alpha,\beta)=\sum_{i=1}^{l}\alpha_i x_i\cdot x_i-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j x_i\cdot x_j \tag{8}$$

由于 β_i 在(8)式中未出现，因而可代之以约束：$0\leqslant\alpha_i\leqslant s_iC, i=1,\cdots,l$，这样优化问题式(3)就转化为对偶问题：

$$\max_{\alpha}\sum_{i=1}^{l}\alpha_i x_i\cdot x_i-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j x_i\cdot x_j$$
$$\text{s.t. } 0\leqslant\alpha_i\leqslant s_iC, i=1,\cdots,l; \tag{9}$$
$$\sum_{i=1}^{l}\alpha_i=1$$

其 KKT 条件为：

$$\alpha_i(R^2+\xi_i-\|x_i-a\|^2)=0, \tag{10}$$

$$\beta_i\xi_i=0. \tag{11}$$

由(9)式可解出 α_i，再由(6)式可知，球心 a 为 $\{x_1,\cdots,x_l\}$ 的加权平均，其系数 α_i。值得注意的是，由(9)式解出的 α_i 只有部分(通常是少部分)大于零，其它均为零。对于那些 $\alpha_i>0$ 所对应的 x_i 称为支持向量。支持向量可分为两类，当 $\alpha_i=s_iC$ 时，由 KKT 条件(10)和(11)式可知，$R^2+\xi_i=\|x_i-a\|^2, \xi_i\geqslant 0$，因而 $\|x_i-a\|^2\geqslant R^2$。此时对应的数据点 x_i 在超球面外部(可能在球面上)。当 $0<\alpha_i<s_iC$ 时，由 KKT 条件可知 $\xi_i=0$，此时对应的数据点 x_i 位于超球面上。

我们把位于上述超球面外部的点视为异常值。一个数据集中异常值的比例与惩罚因子 C 有关，C 的取值较小，则位于超球面外部的数据比例较大；C 的取值较大，则位于超球面外部的数据比例较小。容易看出，当 $\sigma C\geqslant 1$ 时，将没有异常值。事实上，如果 x_i 为异常值，则 $\alpha_i=s_iC>1$ 与 $\sum\alpha_i=1$ 矛盾。

对于一个测试数据 x，为了判别它是否在超球面(由训练数据确定)内部，我们需要计算该数据到球心的距离。当此距离不超过半径 R，即

$$\|x-\sum_{i=1}^{l}\alpha_i x_i\|\leqslant R^2 \tag{12}$$

时接受 x。其中 R 为超球面的半径，其计算是通过求超球面上的点($0<\alpha_i<s_iC$)到球心的距离而得到。比如设 x_{i0} 为位于超球面上的点，则

$$R^2=\|x_{i0}-\sum_{i=1}^{l}\alpha_i x_i\|^2=x_{i0}\cdot x_{i0}-2\sum_{i=1}^{l}\alpha_i(x_{i0}\cdot x_i)+\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j(x_i\cdot x_j) \tag{13}$$

将(13)式代入(12)式得

$$x\cdot x-2\sum_{i=1}^{l}\alpha_i(x\cdot x_i)\leqslant x_{i0}\cdot x_{i0}-2\sum_{i=1}^{l}\alpha_i(x_{i0}\cdot x_i) \tag{14}$$

3. 高斯核方法

我们前面的方法是在样本数据所在的空间中寻求包含有(大多数)数据的超球面。由于数据的分布不一定呈现对称球型，因而不能期望得到一个紧的数据域描述。幸运的是，我们的方法仅涉及数据所在空间的内积，类似于文[2]的方法，我们可以用一个核函数 $K(\cdot,\cdot)$ 代替上面的内积。这意味着将数据映射到一个高维特征空间，但不必知道这个具体的映射。由于多项式核不适用数据域描述[4]，这里我们用高斯核函数：$k(x,y)=\exp(-q\|x-y\|^2)$。此时问题(9)就为

$$\max_{\alpha}\ 1-\sum_{i=1}^{l}\alpha_i^2-\sum_{i\neq j}\alpha_i\alpha_j K(x_i\cdot x_j)$$
$$\text{s.t. } 0\leqslant\alpha_i\leqslant s_iC, i=1,\cdots,l; \tag{15}$$
$$\sum_{i=1}^{l}\alpha_i=1.$$

对于测试数据 x，接受域(14)式为

$$\sum_{i=1}^{l}\alpha_i K(x\cdot x_i)\geqslant\sum_{i=1}^{l}\alpha_i K(x_{i0}\cdot x_i) \tag{16}$$

由此可见，从训练数据到测试数据我们只涉及核函数 $K(\cdot,\cdot)$ 而与具体映射无关。

在具体应用时，需要进一步确定的参数有 σ, s_i, C, q，其中隶属度 s_i 及下界 σ 的确定依赖于具体的训练数据的特征。一般而言，首先必须确定隶属度的下界 σ，然后根据训练数据的特征，如时间特征，权重特征等建立 s_i 与 x_i 之间的联系，具体办法参见文[6]，无论如何 s_i 的确定多多少少带有主观色彩。

惩罚因子 C 的作用在前面已讨论，高斯核函数中的尺度参数 q 可以控制支持向量的个数，当支持向量的个数太多时，应使得 q 取小一些；当支持向量的个数太少时，应使得 q 取大一些。

下面我们考察一个数值例子。假设有一组与时间相关的数据

$$(x_1,s_1,t_1),(x_2,s_2,t_2),\cdots,(x_l,s_l,t_l)$$

其中 $t_1\leqslant t_2\leqslant\cdots t_l$ 是相应数据收集的时间，且时间越靠后，数据越重要。这时可取 $s_i=f(t_i)$ 为 t_i 的单调增加函数且 $s_1=\sigma>0, s_l=1$，比如

$$f(t)=\frac{1-t_1}{t_l-t_1}t+\frac{t_l s_1-t_1}{t_l-t_1}$$

图1给出了一个具体算例。图中每个小圆圈代表一个二维数据点，且圆圈的大小反映了相应的模糊隶属度 s_i 的大小，圆圈越大表示 s_i 越大，大的封闭曲线是由(16)式决定的决策函数，其中实线是使用模糊支持向量数据域描述得到的结果。虚线是不用模糊隶属度(即 $s_i=1$)而得到的结果，从图中可以看出，不用模糊隶属度时，有可能将非常重要的数据视为异常值。

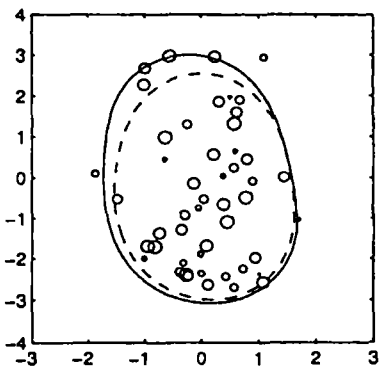

图1

参 考 文 献

1 Vapnik V. Statistical Learning Theory. Wiley, 1998
2 Cristianini N, Shaw-Taylar J. An Introduction to Support Vector Machines. Cambridge, UK: Cambridge University Press, 2000
3 Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121～167
4 Tax D M J, Duin R P W. Support Vector Domain Description. Pattern Recognition Letters, 1999, 20(11-13): 1191～1199
5 Malyscheff A M, Trafalis T B, Raman S. From Support Vector Machine Learning to the Determination of the Minimum Enclosing Zone. Computers and Industrial Engineering, 2002, 42(1): 59
6 Lin C-F, Wang S-D. Fuzzy Support Vector Machines. IEEE Trans. on Neural Networks, 2002, 13(2): 464～471