

支持向量域多分类器

吴德, 刘三阳

(西安电子科技大学计算机学院, 710071, 西安)

摘要: 为解决多分类支持向量机计算量大、训练时间长的问題,构造了支持向量域多分类器(MSVDC).在训练阶段,运用支持向量域描述求得各类样本的最小包围超球,进而将数据空间划分为不同区域;在测试阶段,计算待识别样本与最小包围超球球心的距离,并判断其空间位置;对超球重叠以及超球外区域的样本,定义一种相对类距离,判断样本归属该值较小的类. MSVDC 避免了重复利用训练样本,降低了内存占用并提高了计算效率.数值实验结果表明:MSVDC 具有好的鲁棒性,分类精度可高达 98.89%,分别比一对多和一对一算法高 4.51%和 1.24%,训练时间分别为一对多和一对一算法的 18.06%和 55.41%.

关键词: 多分类器;支持向量域描述;最小包围超球;相对类距离;空间位置

中图分类号: TP301 **文献标志码:** A **文章编号:** 0253-987X(2012)06-0087-05

Multiple Support Vector Domain Classifier

WU De, LIU Sanyang

(School of Computer Sciences, Xidian University, Xi'an 710071, China)

Abstract: To solve the difficulties in multi SVM, such as huge computation and long training time, a multiple support vector domain classifier (MSVDC) is proposed. In the training process, the support vector domain description (SVDD) is employed to obtain the minimal enclosing ball (MEB) of each class, and then the data space is divided into different regions. In the test phase, the distances from the test sample to the MEB centers are evaluated, and the position of the test sample is determined. For samples in the overlapped and outside regions of the MEB, a relative class distance is defined, and is erected in the class with the smallest value. MSVDC avoids the repeated usage of training data, and reduces the memory and enhance the efficiency. Numerical experiments show that MSVDC is endowed with better robustness. The classification accuracy gets to 98.89%, 4.51% and 1.24% higher than “one-against-all” and “one-against-one”, and the training time is only spent for 18.06% and 55.41% of “one-against-all” and “one-against-one”, respectively.

Keywords: multiple classifier; support vector domain description; minimal enclosing ball; relative class distance; space position

由 Vapnik 等提出的支持向量机具有良好的推广性和较好的分类精度,能有效地解决过学习问题,从而被广泛应用在模式识别、函数逼近、数据挖掘等

领域^[1-2].传统支持向量机(SVM)主要针对两类分类问题,而在实际应用中,如数据挖掘、文本分类等需处理的数据是海量和多类的.如何将 SVM 推广,

收稿日期: 2011-10-14. 作者简介: 吴德(1979—),男,博士生;刘三阳(通信作者),男,教授,博士生导师. 基金项目: 国家自然科学基金资助项目(60974082);陕西省教育厅自然科学基金资助项目(2010JK773);西安石油大学博士专项科研基金资助项目(Z10027).

网络出版时间: 2012-02-22

网络出版地址: <http://www.cnki.net/kcms/detail/61.1069.T.20120222.1012.006.html>

以解决多分类问题,是近几年来研究的重点之一。

已有学者提出一对一、一对多支持向量机算法^[1,3-4]。一对一算法在训练每个 SVM 时,视其中某类为正类,其余类为负类,构造 $k(k-1)/2$ 个子分类器,在测试阶段必须计算所有子分类判决函数,计算量大。一对多算法依次选出一类样本与其余类构造分类器,构造 k 个子分类器,通过投票法决定样本的最终归属,但正负类数据分布不平衡,导致分类精度降低。Tax 等借鉴 SVM 思想,提出的支持向量域描述(SVDD)显示出极大的优越性^[5-6],并被广泛应用于单值分类、奇异值监测和雷达目标识别等领域。对 SVDD 的研究包括提取边界向量构造快速学习算法^[7],构造分类器(SVDC)^[8]以及构造多分类算法^[9]。文献[10]提出空间支持向量域分类器(SSVDC),根据样本的空间分布设定相应的分类准则,算法具有较高的分类精度和较短的分类时间。

受文献[10]启发,本文构建了一种支持向量域多分类器(MSVDC),对不同位置的样本设定相应判别准则。首先,运用 SVDD 构造各类样本的最小包围超球,将数据空间划分成单一超球内区域、超球重叠区域和超球外区域;然后,计算待识别样本到最小包围超球球心的距离,根据其半径的大小确定空间位置。对位于某单一超球内的样本,采取 SVDD 的判决准则;对位于超球交叉区域和超球以外区域的样本,以其至各类超球球心的距离与相应超球半径的比值,来定义一种相对类距离,判断样本归属相对类距离的值较小的类。

1 支持向量描述和支持向量域分类器

给定 n 维空间 $\mathbf{R}^n$ 中包含 l 个对象的数据集 $\{\mathbf{x}_i\}_{i=1}^l$,SVDD 致力于寻找一个体积最小的超球,使全部数据 $\mathbf{x}_i$ 均被包围在该超球内。为减少奇异值的影响,对位于超球外的样本引入松弛变量 $0 \leq \xi_i \in \mathbf{R}^1$,并对违反约束的松弛总量加入惩罚 $0 < C \in \mathbf{R}^1$ 。为了得到更为精确的数据描述模型,SVDD 采用一个非线性映射 $\phi: \mathbf{x} \rightarrow \phi(\mathbf{x})$ 将原始空间的数据映射到一个高维的特征空间,并运用核函数技术以实现非线性变换后的样本描述且不增加计算复杂度。记最小包围超球的半径为 R ,球心位置为 $\mathbf{a}$,则训练 SVDD 即为求解如下凸二次规划

$$\begin{aligned} \min \quad & R^2 + C \sum_{i=1}^l \xi_i \\ \text{s. t.} \quad & (\phi(\mathbf{x}_i) - \mathbf{a})^T (\phi(\mathbf{x}_i) - \mathbf{a}) \leq \\ & R^2 + \xi_i (i = 1, \dots, l) \end{aligned}$$

$$0 \leq \xi_i (i = 1, \dots, l)$$

上述规划的最优解可通过 Lagrange 法求解如下对偶规划得到(具体步骤略,请参阅文献[10])

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

式中: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ 为核函数。常用的核函数有线性核函数 $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$, 多项式核函数 $K(\mathbf{x}, \mathbf{y}) = [(\mathbf{x}^T \mathbf{y}) + 1]^q$ 和径向基核函数 $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2)$ 。

记对偶规划的最优解为 $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_l^*)^T$, 称对应 $\alpha_i^* = 0$ 的样本为非支持向量, 称对应 $\alpha_i^* > 0$ 的样本为支持向量 S 。记支持向量的数目为 n_s , 最小包围超球的球心和半径分别为 $\mathbf{a} = \sum_{i=1}^l \alpha_i^* \phi(\mathbf{x}_i)$ 和 $R = 1/n_s \left(\sum_{i \in S} \|\phi(\mathbf{x}_i) - \mathbf{a}\| \right)$ 。

为了将 SVDD 推广到两分类, 仅对正类 $y_i = 1$ 样本建立超球, 而不考虑负类 $y_i = -1$ 的样本^[7]。根据最小包围超球的球心位置 $\mathbf{a}$ 和半径 R , 通过描述边界设计出支持向量域分类器

$$\begin{aligned} f(\mathbf{x}) = \operatorname{sgn} \left\{ R^2 - K(\mathbf{x}, \mathbf{x}) + 2 \sum_i \alpha_i^* y_i K(\mathbf{x}, \mathbf{x}_i) - \right. \\ \left. \sum_{i,j} \alpha_i^* \alpha_j^* y_i y_j K(\mathbf{x}, \mathbf{x}_i) \right\} \\ R^2 = \frac{1}{n_s} \sum_{i \in S, y_i = 1} \left\{ K(\mathbf{x}, \mathbf{x}_i) - 2 \sum_i \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + \right. \\ \left. \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right\} \end{aligned}$$

2 多分类器 MSVDC

已有的一对一、一对多支持向量机, 通过构造多个 SVM 分类器并组合来实现, 但存在计算量大、训练时间长、不容易判断交叉重叠区域样本的类别等问题。参照文献[10]的思想, 本节构造支持向量域多分类器 MSVDC, 根据样本的空间分布设定相应的分类准则。训练时, 通过 SVDD 的描述边界对数据空间进行分割, 并对不同子空间设置对应的分类准则; 测试时, 判断待识别样本所处位置, 并采取相应分类准则。

给定待测样本 $\mathbf{z}$, 计算其到第 i 类样本点的超球球心 $\mathbf{a}_i$ 的距离 $\|\mathbf{z} - \mathbf{a}_i\|$ 与超球半径 R_i 的比值, 并定义该值为 $\mathbf{z}$ 的相对类距离 $d_i = \|\mathbf{z} - \mathbf{a}_i\| / R_i$ 。3 条空间分类准则如下。

(1)对位于某单一超球类内的样本,判断样本归属该超球对应的类,此时第 i 类样本的相对类距离 d_i 满足 $d_i \leq 1, d_j > 1 (i \neq j)$.

(2)对位于超球重叠区域的样本, $d_i \leq 1 (i = 1, \dots, c; c \geq 2)$,即有多于两个的相对类距离不大于1,判断样本归属该值最小的类.

(3)对位于超球外区域的样本, $d_i > 1 (i = 1, 2, \dots, k)$,即所有的相对类距离大于1,判断样本归属该值最小的类.

如果出现几个相同的 d_i 而分母 R_i 不同,则比较分子 $\|z - a_i\|$ 的大小,并判断样本归属分子之值最小的类.如果出现几个相同的 d_i 且分母 R_i 亦相同,说明这些类的分布相似,而且待测样本处于对称点,此时判断样本归属任一类都是正确的.

以二维空间的三分类为例,直观表述本文算法.在训练阶段,MSVDC将数据划分为单一超球类内区域A、B、C,超球重叠区域(两个及两个以上超球的重叠区域)D和超球外区域E,如图1所示.在测试阶段,MSVDC判断待测样本所处区域,并采取相应分类准则.

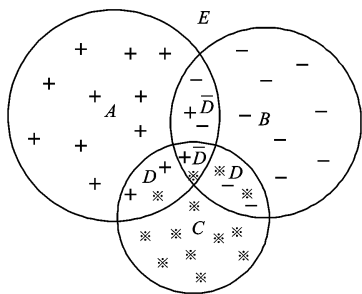


图1 MSVDC的分类原理示意图

有了以上分析,得到基于支持向量描述的空间分类器MSVDC如下.

步骤0:输入 k 类训练样本 $\{x_i, y_i\}_{i=1}^l (x_i \in \mathbf{R}^n, y_i \in \{1, 2, \dots, k\})$ 、惩罚因子和核参数的初始值 $C_1, C_2, \dots, C_k$ 和 $\sigma_1, \sigma_2, \dots, \sigma_k$.

步骤1:构造对应 $y_i = 1, y_i = 2, \dots, y_i = k$ 的 k 类样本的最小包围超球,记作 $(a_1, R_1), (a_2, R_2), \dots, (a_k, R_k)$.

步骤2:计算待测样本 z 到球心 $a_1, a_2, \dots, a_k$ 的距离 $\|z - a_1\|, \|z - a_2\|, \dots, \|z - a_k\|$.

步骤3:计算 k 个相对类距离的值, $d_i = \|z - a_i\| / R_i (i = 1, 2, \dots, k)$.

步骤4:根据3条分类准则,判断待识别样本的归属.

由于MSVDC仅对单一类样本运行SVDD,处

理的样本规模较小,预计训练时间较短.由于MSVDC充分利用每类样本的信息,预计分类精度较高.由于MSVDC根据样本的空间位置设定分类准则,预计鲁棒性较好,对参数变化不敏感.对位于单一类内的样本,判断归属该类;对位于类交叉区域及类外的样本,采用相对类距离判别.

给定 l 个 k 类训练样本,设每类有 l/k 个,对比MSVDC与一对多和一对一算法的运算量. MSVDC构造 k 个超球作为子分类器,每个子分类器仅需 l/k 个样本参与训练.一对多算法构造 k 个子分类器,每个子分类器需要全部 l 个样本参与训练.一对一算法构造 $k(k-1)/2$ 个分类器,每个子分类器需要 $2l/k$ 个样本参与训练.后两种算法都重复利用了训练样本,复杂度较高,训练时间较长.显然,MSVDC和一对多算法构造的子分类器个数远远小于一对一算法的分类器个数 $k \leq k(k-1)/2$,且MSVDC训练子分类器时参与训练的样本数远远小于一对一和一对多算法,即 $l/k \leq (2l/k) \leq l$.

3 数值实验及结果分析

本文实验均在DELL PC机(CPU为P4, 3.06 GHz,内存为0.99 GB)上利用Matlab 7.4编程实现.实验采用性能良好的高斯核函数 $K(x, y) = \exp(-\|x - y\|^2 / \sigma^2)$.首先,对小规模的正态分布数据进行实验测试,并研究不同核参数下多分类器MSVDC的分类表现;其次,对大规模的正态分布数据和Yale Face数据,对比一对一和一对多分类算法的数值实验结果.

例1:产生均值和方差都相同的3类正态分布数据180个,其中每类各60个.随机选取其中90个作为训练样本,其余作为测试样本,为避免随机性,取10次实验的平均结果.设定惩罚因子 $C=1$,MSVDC的分类性能见表1.

从表1可以得出如下结论.

(1)MSVDC具有好的鲁棒性.当核参数取 $\sigma = 0.04 \sim 0.5$ 时,MSVDC的整体训练和测试精度受核参数 σ 变化的影响不大.

(2)MSVDC的分类精度很高.训练精度高达100%,测试精度高达98.89%,结果令人满意.

(3)MSVDC的训练和测试时间很短.对于不同核参数的取值,MSVDC的训练时间大多为0.26 s,测试时间在0.46 s左右.

需要指出的是:虽然对不同的核参数取值,MSVDC对3类样本都具有较高的训练精度(在

96.67%~100%之间),但过低的核参数会导致对训练样本的过拟合,使得测试精度较低,如 $\sigma=0.015$ 时,对第二类的测试精度仅为40%。实际运用时,可以通过参数空间穷尽搜索法得到最优参数。

例 2: 产生均值和方差均不相同的 3 类正态分

布数据 1 800 个,其中每类各占 600 个,随机选取其中 900 个作为训练样本,其余作为测试样本. 设定惩罚因子 $C=1$,采用十重交叉验证法选取最优核参数,并列出不同算法对 10 次随机抽取实验的分类性能见表 2。

表 1 不同核参数下分类性能比较

σ	训练精度/%				测试精度/%				支持向量数			整体训	整体测
	第一类	第二类	第三类	整体	第一类	第二类	第三类	整体	第一类	第二类	第三类	练时间/s	试时间/s
0.015	100.00	100.00	100.00	100.00	63.33	40.00	100.00	67.78	30	30	30	0.26	0.58
0.030	100.00	100.00	100.00	100.00	93.33	66.67	96.67	85.56	28	27	27	0.28	0.46
0.040	100.00	100.00	100.00	100.00	100.00	86.67	96.67	94.44	26	26	25	0.26	0.50
0.100	100.00	100.00	100.00	100.00	100.00	100.00	90.00	96.67	17	16	16	0.26	0.43
0.200	100.00	100.00	100.00	100.00	100.00	100.00	96.67	98.89	8	6	8	0.29	0.40
0.500	100.00	96.67	100.00	98.89	100.00	100.00	96.67	98.89	4	2	3	0.28	0.37

表 2 不同算法的分类性能比较

算法	训练	测试	支持向	训练	测试
	时间/s	时间/s	量数	精度/%	精度/%
MSVDC	70.32	89.48	13	98.78	98.89
一对一	126.91	93.27	43	99.03	97.65
一对多	389.26	126.91	96	97.91	96.27

注:对应 MSVDC 的训练精度和测试精度是第一、第二和第三类精度的平均值,这 3 类样本的训练精度分别为 98.00、99.00 和 99.33,测试精度分别为 99.00、99.67 和 98.00。

根据表 2 数据,得到以下结论。

(1)MSVDC 在最优参数下的训练和测试精度均很高。MSVDC 具有与一对一算法接近的结果,且均优于一对多算法。这是由于一对多算法每次训练时都用到所有的训练样本,构造的分类面比较复杂。

(2)MSVDC 具有最少的支持向量数、最短的训练和测试时间。由于每次处理的样本规模较小,MSVDC 的支持向量数、训练和测试时间远低于一对一和一对多算法。一对多算法的支持向量数最多,导致其测试时间也最长。

例 3:多分类 Yale Face 数据. 该数据库包含 15 个人不同表情的灰度图像,每人 11 幅,图像分辨率为 320×243 。实验前通过剪切将图像分辨率降为 21×30 ,并采用 PCA 降维,依据能量损失不超过 15% 的原则抽取 20 个主成分。依次按照训练集与测试集的比例分别为 4 : 7、5 : 6、6 : 5 和 7 : 4 选取每个个体的若干样本参与训练。对比本文算法和 SVM 的结果见表 3。

表 3 不同算法的精度和时间比较

算法	精度/%				时间/s			
	4 : 7	5 : 6	6 : 5	7 : 4	4 : 7	5 : 6	6 : 5	7 : 4
一对多	88.92	89.33	90.54	91.71	10.13	15.61	24.52	35.28
一对一	90.41	91.25*	91.63*	92.18	3.05	4.17	6.31	8.09
MSVDC	90.53*	90.37	90.41	92.85*	2.13*	2.27*	2.38*	2.41*

注:加 * 数字代表最高的分类精度和最短的训练时间。

由表 3 得知:

(1)MSVDC 具有与一对多和一对一算法相当 的分类精度;

(2)MSVDC 具有最短的训练时间,随训练数据的增多,一对多和一对一算法的训练时间迅速增加,而 MSVDC 没有重复使用训练样本,训练时间基本不变。

4 结 论

传统的一对一支持向量机算法需构造的子分类器数目随类别数成超线性增长,且测试阶段须计算所有子分类函数,计算量大。一对多支持向量机算法依次选出一类样本与其余类构造分类器,正负类数据分布不平衡,分类精度较低。MSVDC 运用 SVDD 对各类样本构造 MEB,对完全位于单一超球内的样本,判断该样本归属相应超球,对位于超球交叉区域和描述边界以外的样本,定义一种相对类距离,根据其值判断样本类别。正态分布数据集上的数值实验表明,MSVDC具有较好的稳定性和较高的分类精度。

参考文献:

- [1] 许建华. 统计学习理论[M]. 张学工,译. 北京:电子工业出版社,2004: 353-503.
- [2] VAPNIK V N. An overview of statistical learning theory[J]. IEEE Trans on NN, 1999, 10(3): 988-999.
- [3] 唐发明,王仲东,陈绵云. 支持向量机多类分类算法研究[J]. 控制与决策,2005,20(7):746-754.
TANG Faming, WANG Zhongdong, CHEN Mi-anyun. On multiclass classification methods for support vector machines [J]. Control and Decision, 2005, 20(7): 746-754.
- [4] HSU L C. A Comparison of methods for multiclass support vector machines[J]. IEEE Trans on Neural Networks, 2002, 13(2): 415-425.
- [5] TAX D M J, DUIN R P W. Support vector data description [J]. Machine Learning, 2004, 54(1): 45-66.
- [6] 冯国瑜,肖怀铁,付强,等. 基于自适应SVDD的雷达目标分类方法[J]. 系统工程与电子技术,2011,33(2): 253-258.
FENG Guoyu, XIAO Huaitie, FU Qiang, et al. Method of radar target classification based on adaptive SVDD [J]. Systems Engineering and Electronics, 2011, 33(2): 253-258.
- [7] LIANG Jinjin, LIU Sanyang, WU De. Fast training of SVDD by extracting boundary targets [J]. Iranian Journal of Electrical and Computer Engineering, 2009, 8(2):133-137.
- [8] 陆从德,张太镒,胡金燕. 基于乘性规则的支持向量域分类器[J]. 计算机学报,2004,27(5):690-694.
LU Congde, ZHANG Taiyi, HU Jinyan. Support vector domain classifier based on multiplicative updates [J]. Chinese Journal of Computers, 2004, 27(5): 690-694.
- [9] ZHUANG Jinfa, LUO Jian, PENG Yanqing, et al. Rejected transductive inference multi-class SVDD[J]. Journal of Information and Computational Science, 2009, 6(2): 829-836.
- [10] 梁锦锦,刘三阳,吴德. 空间支持向量域分类器[J]. 西安电子科技大学学报,2008,35(6):1080-1083.
LIANG Jinjin, LIU Sanyang, WU De. Space support vector domain classifier [J]. Journal of Xidian University,2008,35(6):1080-1083.

(编辑 杜秀杰)