DOI:CNKI:50-1165/N.20121112.1642.012

基于多核函数的模糊支持向量机学习算法*

徐国浪[1],魏 延[2]
(1. 重庆师范大学 数学学院;2. 计算机与信息科学学院,重庆 401331)

摘要:作者针对单个核函数构成的 SVM 并不能满足诸如数据异构或不规则、样本规模巨大、样本分布不平坦等实际应用的需求,而将多个核函数进行组合,以获得更好的效果,提出一种基于多核的模糊支持向量机算法。此算法决策树中的模糊核权重主要是借助于样本各自的模糊因子来确定。仿真实验数据表明:与传统单核函数支持向量机相比,多核模糊支持向量机具有很好的优越性。

关键词:多核;模糊集;模糊支持向量机;多核分类算法
中图分类号:TP38 **文献标志码**:A **文章编号**:1672-6693(2012)06-0050-04

支持向量机(Support vector machine, SVM)是建立在统计学习理论基础上的一种核机器学习方法,由 Vanpink 等人于 1995 年提出[1-2],它在解决有限样本、非线性及高维模式识别问题中表现出许多特有的优势。目前,已经广泛应用到文本分类、人脸检测、生物信息和入侵检测技术等多个领域[3-7],是解决分类和回归问题的有力工具之一。

SVM 能够较好地解决小样本、非线性和高维模式识别等实际问题,有较强的泛化能力。SVM 采用的核函数将输入空间中的样本映射到某一高维特征,将原空间中的非线性问题转化为高维特征空间中的线性问题,从而在高维特征空间内求得最优分类面。因此,在解决非线性分类或回归问题过程中,核函数的选取非常重要。然而,对于实际问题,由于各个分类对象之间差异较大,因此想较为容易地找到一个合适的核函数对分类样本进行正确分类比较困难,往往是靠大量实验经验所得来完成。因此,本文提出组合多种核函数的支持向量,利用不同核函数之间具有互补性特点,构建多核支持向量机。

由于一般支持向量机在训练的时候对所有训练点同等对待,将每一训练点的全部信息加以学习,这就产生一定的局限性。例如,由于样本的每一个训练点起的作用不同,支持向量起决定性作用,非支持向量基本不起作用,而噪音点或野值点则对正确分类起负面作用,所以分类时应对不同的训练点加以区别对待,尽可能保持支持向量,剔除非支持向量,消除噪音点和野值点[8]。为了满足上面要求,Lin[9-11]等建了模糊支持向量机(Fuzzy support vector machine,FSVM),正是由于模糊支持向量机具有比一般的支持向量机更好的特性,故本文选择在模糊支持向量机基础之上研究多核支持向量机。

1 模糊支持向量机

在传统的单核支持向量机中,最优分类面常常由少量的位于分类边缘的支持向量来决定,而含野值点的样本往往也位于边缘,所以支持向量机在训练过程中对于外围野值点十分敏感,为了减少传统支持向量中异常数据点对支持向量机的影响,Lin 等构建了模糊支持向量机[10]。模糊支持向量机可以减小甚至忽略非重要样本和噪声点对支持向量机学习的影响,从而提高分类精度。

定义 1 设 X 是一个非空集合,则称

$$F=\{\langle x,u_F(x_i)\rangle \mid x\in X,i=1,2,\cdots,l\} \quad (1)$$

为模糊集,$u_F(x_i)$ 为样本 x 中第 i 个属于 F 模糊集的隶属度,且 $u_F(x_i)$ 在[0,1]取值。

在采用模糊支持向量机时,首先要对数据进行预处理:选择一个合适的隶属度函数,得到每一个样本 x_i 的隶属度函数 $u_F(x_i)$。于是便可以得到新的模糊训练集 $\{x_1,y_1,u_F(x_i)\}$,$i=1,2,\cdots,l$。每个训练点 $x_i\in R^d$,$y_i\in\{-1,+1\}$,$0\leqslant u_F(x_i)\leqslant 1$。其中 $u_F(x_i)$ 为训练集 $\{x_1,y_1,u_F(x)\}$ 的输出 $y_i=1$(正类)或 $y_i=-1$(负类)的模糊隶属度($i=1,2,\cdots,l$)。

模糊隶属度 $u_F(x_i)$ 是指训练集 $\{x_1,y_1,u_F(x)\}$ 隶属某一类的程度,而 ξ_i 是对错分程度的度量,所以就用 $u_F(x_i)\cdot\xi_i(i=1,2,\cdots,l)$ 衡量对于重要程度不同的变量错分程度。由此得到最优分类超平面

* 收稿日期:2012-04-10 网络出版时间:2012-11-12 16:42:01
资助项目:重庆市教委科技计划项目(No. KJ090823);重庆师范大学博士研究基金(No. 11XLB047)
作者简介:徐国浪,硕士研究生,研究方向为机器学习与智能计算;通讯作者:魏延,E-mail:weiyan@cqnu.edu.cn
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20121112.1642.201206.50_012.html

的目标函数的最优结构

$$\min \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} u_F(x_i)\xi_i$$
$$\text{s.t.} \quad y_i[\omega\cdot\varphi(x_i)+b]-1+\xi_i\geqslant 0 \qquad (2)$$
$$\xi_i\geqslant 0, i=1,2,3,\cdots,l$$

其中,惩罚因子 C 为常量,ω 表示线性分类函数的权系数,$\xi=(\xi_1,\xi_2,\cdots,\xi_l)^{\mathrm{T}}$,$\varphi(x_i)$ 是将 x_i 从 R^d 映射到高维特征空间。而相应的最优分类超平面的判别函数式为

$$f(x) = \operatorname{sign}\Big[\sum_{i=1}^{l}\alpha_i^* y_i K(x_i,x) + b^*\Big] \qquad (3)$$
$$0\leqslant\alpha_i^*\leqslant\xi_i C, i=1,2,\cdots,l$$

其中,$K(x_i,x)$ 为核函数,其作用就是将高维特征空间的内积运算转化为低维模式空间上的一个简单的函数计算。

由(2)式可以看出,当 $u_F(x_i)$ 的值很小时,减少了 ξ_i 在式中的影响,以至于将相应的 x_i 可以视为不重要的样本。对于野值点或噪音样本,如果能使其 $u_F(x_i)$ 很小,则此样本对支持向量机的训练作用大为降低,进而降低了它们对训练支持向量机的影响。由此可见模糊因子 $u_F(x_i)$ 的确定正是这种模糊支持向量机工作性能好坏的关键[12]。

目前常用的核函数主要有 3 类:多项式核函数(Poly)、径向基核函数(Rbf)、Sigmoid 核函数。

2 模糊隶属度函数的确定

模糊隶属度函数的确立方法很多,但其常见的基本思想有以下两类:一类是基于类中心距的隶属度函数设计方法;另一类是基于密度法的隶属度函数设计思想。前者的优点是简单易行,但是其类中心半径的确定对噪声和孤立点很敏感。后者准确率比较高,然而其复杂度较高,在相同条件下,计算量明显增加。本文采用 C-均值聚类与密度法相结合的双隶属法来确定模糊隶属度函数 u_i。

根据样本点到类中心的距离将样本分为两类,分别采用不同的隶属度方法[12]。

第 1 类 类中心附近的样本点的隶属度。类中心附近的点是指符合下列条件的样本

$$D(x_i,o^+)\leqslant\beta R^+, y_i=+1$$
或
$$D(x_i,o^-)\leqslant\beta R^-, y_i=-1$$

其中 $0<\beta\leqslant 1$ 为可调参数。则其隶属度函数为

$$u_i=\begin{cases}\dfrac{1}{1+\dfrac{\gamma D(x_i,o^+)}{R^+}}, y_i=+1\\[2ex]\dfrac{1}{1+\dfrac{\gamma D(x_i,o^-)}{R^-}}, y_i=-1\end{cases} \qquad (4)$$

其中 $\gamma>0$ 为可调参数。

第 2 类:远离中心的样本点隶属度。远离类中心的样本点是指符合下列条件的点

$$D(x_i,o^+)>\beta R^+, y_i=+1$$
或
$$D(x_i,o^-)>\beta R^-, y_i=-1$$

其中 $0<\beta<1$ 为可调参数。则其隶属度函数为

$$u_i=\frac{\eta\rho^+(x_i,\theta)}{\rho^+(x_i,\theta)+\rho^-(x_i,\theta)} \qquad (5)$$

其中 $0<\eta\leqslant 1$,$\theta>0$ 为可调参数。

3 多核模糊支持向量机及其算法

虽说模糊支持向量机比一般的支持向量机具有更佳的优越的特性,但是仍然存在以下两方面的缺陷:1)像(3)式的决策树只能对应于某一类特殊的函数集,而不是一组混合的函数集。如核函数取多项式核函数,则对应的就是一组多项式函数集,而且不能是多项式和径向基函数集的混合集;2)大部分核函数只能有一个自由参数来控制其推广性能(如径向基核函数中由宽度参数来控制),不能同时用多个不同参数。对于现实生活中分类问题,核函数的选择至关重要,直接关系到分类正确率。通常采用穷举不同的核函数来获得多个分类正确率,取正确率最高的函数作为该函数的实际问题的核函数。当数据异构或不规则、样本规模巨大、样本不平坦分布时,这种方法实现起来很麻烦、很费时,其精度也会因此下降。那是否有一种方法能解决这种问题呢[13]?本文就探索一种组合的多核模糊支持向量机算法,其决策树和算法步骤如下。

决策树

$$f(x) = \operatorname{sign}\Big\{\sum_{i=1}^{l}\alpha^* y_i \hat{K}(x_i,x) + b^*\Big\} \qquad (6)$$

其中,$\hat{K}(x_i,x)=\sum_{i=1}^{l}\sum_{j=1}^{3}u_i K_j(x_i,x)$,$u_i$ 表示样本集 x 中第 i 个样本的隶属度,在这里 u_i 作为多核支持向量机的权重。$K_j(x_i,x)$ 表示第 j 类核函数,在本文中 j 取 1,2,3 时,分别对应常见的 Poly、Rbf 和 Sigmoid 3 类核函数。

在描述模糊多核函数之前,首先简要回顾一下 Mercer 定理,它是判定核函数的充要条件。令 X 是 $\mathbf{R}^n$ 的紧子集。假定 K 是连续对称函数,存在积分算子 $T_K: L_2(X)\to L_2(Y)$,使得$(T_K f)(\cdot)=\int_X K(\cdot,u)f(u)\,\mathrm{d}u$ 是正的,也就是 $\int_{X\times X}K(u,v)f(v)\,\mathrm{d}u\mathrm{d}v\geqslant 0$,$\forall f\in L_2(X)$ 扩展 $K(u,v)$ 到一个一致收敛的序列(在 $X\times X$ 上),这个序列由 T_K 的特征函数 $\Phi_i\in L_2(X)$ 构成,归一化使得 $\|\Phi_i\|_{L_2}=1$,并且 $\lambda_i\geqslant 0$,则有

$$K(u,v) = \sum_{i=1}^{\infty}\lambda_i\Phi_i(u)\Phi_i(v) \qquad (7)$$

定理1 如果 K 为核函数,则称 $\hat{K}$ 为多核模糊核函数,$\hat{K}(x_i,x)=q(K(x_i,x),u)$,其中

$$q(K(x_i,x),u)=\sum_{i=1}^{l}\sum_{j=1}^{m}u_iK(x_i,x),\ u_i\geqslant 0$$

证明 固定一个有限点集$\{x_1,x_2,\cdots,x_l\}$,令 K_1 和 K_2 是限制在这些点上的相应的矩阵。考虑任意向量 $u\in\mathbf{R}^l$,矩阵 K 是半正定的充要条件是对于所有的 u_i,$u_1Ku_2\geqslant 0$,因此,$u_1(K_1+K_2)u_2=u_1K_1u_2+u_1K_2u_2\geqslant 0$,这样 K_1+K_2 是半正定的,即 K_1+K_2 满足 Mercer 条件,所以是核函数[14]。 证毕

另外,需要说明的是,在选择两个及两个以上核函数时将如何组合的问题:这里分两种情况:1)核函数个数为偶时,先采取两个两个组合,然后逐渐成倍递减构成复合的核函数;2)核函数个数为奇数时,先两个两个组合,组合剩下的单个核函数算着一组,然后如上面1)组合方法一致。以此类推,最终形成基于多核模糊支持向量机接下来,验证定义中的组合核函数是满足 Mercer 性质。

引理1 Mercer 核的非负线性组合仍为 Mercer 核。

证明 $K_i,i=1,\cdots,m$ 是 Mercer 核,令$K(x_i,x)=\sum_{i=1}^{m}\alpha_iK_i(x_i,x)$,其中系数 $\alpha_i\geqslant 0$ 为非负数。因为 K_i 为 Mercer 核,则有

$$\int K_i(x_i,x)f(x_i)f(x)\,\mathrm{d}u\mathrm{d}v\geqslant 0,\ \forall f\in L_2(X)$$

故对于线性组合 $K(x_i,x)$ 有

$$\int K_i(x_i,x)f(x_i)f(x)\,\mathrm{d}u\mathrm{d}v=$$

$$\int\sum_{i=1}^{m}\alpha_iK_i(x_i,x)f(x_i)f(x)\,\mathrm{d}u\mathrm{d}v=$$

$$\sum_{i=1}^{m}\int\alpha_iK_i(x_i,x)f(x_i)f(x)\,\mathrm{d}u\mathrm{d}v\geqslant 0 \qquad (8)$$

对于特殊情况,当 $\sum_{i=1}^{m}\alpha_i=1$ 时 $K(x_i,x)$ 为 $K_i(x_i,x)$ 的凸组合 Mercer 核函数。 证毕

定理2 如果 K 是 Mercer 核,则组合的模糊多核核函数 $\hat{K}(x_i,x)=q(K(x_i,x),u)$ 也为 Mercer 核。

证明 由定理1,已经证出只要 K 是核函数,则 $\hat{K}$ 也是核函数;另外,由引理1,可以证得 Mercer 核的非负线性组合仍为 Mercer 核。故定理2得证。 证毕

由定理2知,Mercer 核的多项式组合仍然为 Mercer 核。因此,以定理2为基础,可以利用现有常用核函数构造一类模糊多核核函数,这类核同时具有平移不变性和旋转不变性,能够适用于各类样本集的学习。此外,使用这类核函数的支持向量机,虽然看起来比较复杂,但是主要工作量是由计算机仿真来完成,给使用者节约许多试验的时间。

算法步骤 Step1 初始化,将原始样本数据进行初步筛选。Step2 将初步筛选后的数据按照(1)式的方法,先建立起模糊集。Step3 根据(4)式和(5)式中方法确定新样本各点模糊隶属度。Step4 根据 $q(K(x_i,x),u)=\sum_{i=1}^{l}\sum_{j=1}^{m}u_iK(x_i,x)$ 选定不同的核函数进行组合。Step5 运用新提出决策树(6)式对多核模糊支持相机进行训练和测试。

4 实例分析

4.1 数据的选取

UCI 数据[15]是机器学习的一个标准数据库,可以用来衡量各种模式下支持向量机算法的有效性。为了验证所提出的基于多核情况下模糊支持向量机的有效性,特选取 UCI 数据库上两种数据共699个样本,每个样本有11个属性,除去16个有未知属性样本外,还有683个:一类 Training,共350个;另一类 Testing,共333个。

4.2 实验结果及其分析

采用前面介绍的多核模糊支持向量机算法,对所选取的683个样本进行试验,选取多项式核函数和径向基核函数进行组合,相关参数设置如下:$C=1,d=1,2,\cdots,5$;$\sigma=0.001,0.010,\cdots,10.000$;$\beta_0=2,\beta_1=-0.8,-0.9,-1.0,-1.2,-1.3$。

采用样本识别的正确率为 R 作为衡量各类分类器的分类精度。

本文的实验是在 MATLAB 7.X 软件上编程实现的,先选取各个单独核函数的 FSVM 获得的正确率,如表1所示;然后采用双核核函数支持向量机对样本进行试验,获得的正确率如表2所示;表3是采用本文提出的方法获得的正确率。

表1 单个核函数支持向量机测试获得正确率 %

序号	核函数类型	最佳准确率	平均准确率
1	Ploy	76.67	65.51
2	Rbf	82.33	66.60
3	Sigmoid	69.34	62.86

由表1中的实验数据可以看出,单个核函数支持向量机获得的平均正确率在65%左右,获得的最佳准确率普遍比平均准确率要高出许多,这就说明要想获得较佳的实验效果,需要以做大量的实验为基础[16-17]。

表2 两种类型核函数支持向量机测试获得正确率 %

序号	核函数组合方式	最佳准确率	平均准确率
1	Ploy 和 Rbf	83.46	76.36
2	Ploy 和 Sigmoid	74.24	70.62
3	Rbf 和 Sigmoid	80.30	68.86

由表 2 的实验数据可以看出:就整体而言,双核支持向量机相比于单核支持向量机,不仅在平均准确率高,而且最佳准确率也相对较高。这说明多核支持向量机要比单核支持向量更具有泛化性。

表 3 采用本文模糊多核核函数支持向量机测试获得正确率 %

序号	核函数组合方式	最佳准确率	平均准确率
1	Ploy、Rbf 和 Sigmoid	87.64	84.67

观察表 3,可以看出:本文提出多核模糊支持向量机,不仅在最佳准确率方面要比单核和双核支持向量机要高,而且在平均准确率方面要远远高于单核支持向量机和双核支持向量机,这也很好地说明本文提出的多核模糊支持向量机在处理结构复杂数据方面更具有泛化性。通过观察 3 个表的纵向数据,不难得出如下结论:多核(包括双核)SVM 比传统的单核 SVM 具有更强的泛化性。

5 结论

本文是在模糊支持相机的基础之上,提出多核的 FSVM 算法,通过严格证明和实验数据证明了此方法在解决具有多种复杂特性样本分类问题上的可行性。这一方法突破了传统单纯依靠通过大量实验来寻找最优核函数的瓶颈。将此多核模糊支持向量机向工程应用方面推广将是笔者未来的研究方向之一。

参考文献:

[1] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):33-42.

[2] 张学工. 模式识别[M]. 北京:清华大学出版社,2009.

[3] Joachims T. Text categorization with Support vector machines:Learning with many relevant features[C]//Proceedings of the European Conference on Machine Learning,Berlin:Springer,1998:137-142.

[4] Waring C A,WLiu X. Facedetection using spectral histograms and SVMs[J]. IEEE Trans Systems,Man and Cybernetics,2005,35(3):467-476.

[5] Guyon I,Weston J,Barnhill S,et al. Gene selection for cancer classi cation using support vector machines[J]. Machine Learning,2002,46(1):389-422.

[6] 饶鲜,董春曦,杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报,2003,14(4):798-803.

[7] 李昆仑,黄厚宽,田盛丰. 模糊多类支持向量机及其在入侵检测中的应用[J]. 计算机学报,2005,28(4):274-279.

[8] 哈明虎. 直觉模糊支持向量机[J]. 河北师范大学学报:自然科学版,2011,31(3):225-229.

[9] LIN C F,WANG S D. Fuzzy support vector machine[J]. IEEE Ransaction on Neural Networks,2002,13:464-471.

[10] LIN C F,WANG S D. Fuzzy support vector machines with automatic membership setting[J]. Studies in Fuzziness and Soft Computing,2005,177:233-254.

[11] Li M J,Ng M K,Cheung Y M,et al. Agglomerative fuzzy K-Means clustering algorithm with selecttion of number of clusters[J]. IEEE Trans Knowledge and Data Engineerning,2008,20:1519-1534.

[12] 邬啸,魏延,吴霞. 改进的双隶属度模糊支持向量[J]. 重庆师范大学学报:自然科学版,2011,28(5):49-52.

[13] 许桂梅,黄圣国. 基于多核支持向量机的飞机重着落诊断[J]. 计算机工程,2011,37(10):157-159.

[14] 邬啸,魏延,吴霞. 基于混合核函数的支持向量机[J]. 重庆理工大学学报:自然科学版,2011,25(10):66-70.

[15] Blake C L,Merz C J. UCI Repository of machine learning databases[EB/OL]. (2012-04-06)[2008-12-12]. http://www.us.uci.edu/vmlearn/mlRepository.html.

[16] 孔浩,杨勇,王国胤. 基于多分类器融合的语音识别方法研究[J]. 重庆邮电大学学报:自然科学版,2011,23(4):492-495.

[17] 袁正午,朱冠宇,丰江帆,等. 基于支持向量机的视频语义场景分割算法研究[J]. 重庆邮电大学学报:自然科学版,2010,22(4):458-463.

Learning Algorithm Based on Fuzzy Support Vector Machine of Multi-Core Functions

XU Guo-lang[1], WEI Yan[2]

(1. School of Mathematics of Chongqing Normal University;
2. College of Computer and Information Science, Chongqing 401331, China)

Abstract: In order to solve those question which a kernel function can not meet, such as heterogeneous or irregular data, large sample size, uneven distribution of the sample the actual application requirements, and obtain better results, the author of this paper proposes a support vector machine algorithm based on multi-core fuzzy. Fuzzy kernel weights of this decision tree algorithm is mainly determined by the fuzzy factors of the sample. The simulation data show that multi-core fuzzy support vector machine has the good the superiority compared with the traditional single-kernel function support vector machine.

Key words: multi-core; fuzzy set; fuzzy support vector machine; multi-core classification algorithm

(责任编辑 游中胜)