混合核函数中权重求解方法①

王行甫, 俞 璐

(中国科学技术大学 计算机科学与技术学院, 合肥 230022)

摘 要: 为了克服支持向量机(SVM)中单核函数的局限性, 经常使用混合核函数做预测, 但混合核函数中各函数权重难以确定. 为解决该问题, 提出了一种基于特征距离的权重求解方法. 该方法首先利用支持向量机的几何意义, 根据同类样本特征距离最小化和异类样本特征距离最大化原理, 分析得出优化函数, 然后对优化函数求解得出权重系数. 实验结果表明, 与传统的交叉验证法和 PSO 算法相比, 该方法在保证预测精度的情况下, 将计算时间减少了 70%左右.

关键词: 支持向量机; 核函数; 权重; 特征距离

Weight Solving Method in Hybrid Kernel Function

WANG Xing-Fu, YU Lu

(School of Computer Science and Technology, University of Science and Technology of China, Hefei 230022, China)

Abstract: In order to overcome the limitation of single kernel in Support Vector Machine(SVM) model, hybrid kernel is usually used in forecasting. However, the weight of functions in the hybrid kernel is hard to calculate. To solve this problem, we propose a new method based on feature-distance. This method firstly gets an optimization function based on SVM's geometric meaning and a principle, which is the feature-distance of the same kind should be minimized and the different should be maximized, and then analyzes the optimization function to work out the weight. Experimental results show that compared with the cross validation method and PSO algorithm, this method reduces the computing time nearly by 70% with the accuracy kept unchanged.

Key words: SVM; kernel function; weight; feature-distance

SVM(Support Vector Machine)是 Vapnik[1]等人于 20 世纪 90 年代提出的基于统计学习理论的机器学习方法, 建立在 VC 维理论和结构风险最小化原则基础之上, SVM 通过将低维空间的输入向量映射到高维特征空间, 实现线性可分, 并利用核函数技巧, 有效的克服了映射函数会引发的维数灾难问题.

在SVM中, 核函数的选择决定了特征空间的结构, 所以核函数直接决定了分类效果的好坏. 目前所使用的核函数基本都是基于单个特征空间的单核函数, 每个核函数的特性并不相同, 所以在不同的应用场景需要采用不同的核函数, 而核函数的构造和选择至今都没有完善的理论依据, 针对一些样本特征维数较高, 采用单核映射无法处理的问题, 近年来出现了多核学习方法[2,3], 将多个核函数进行组合, 以期获得更好的性能.

在多核函数中, 最简单有效的是混合核函数, 虽然混合核函数提高了分类的精度, 但是也引入了权重系数需要确定, 目前系数求解方法大多比较耗时, 也有很多是基于经验值确定, 无法准确求得结果, 本文提出一种基于特征距离求解权系数的方法, 并通过实验验证了方法的高效性, 解决了权系数难以确定的问题.

① 基金项目:国家科技重大专项(2012ZX10004-301-609);国家自然科学基金(61272472,61232018,61202404);安徽省教学研究计划 2010
收稿时间:2014-07-16;收到修改稿时间:2014-09-10

1 核函数

支持向量机中为了使线性不可分的样本变得线性可分, 将原样本空间通过函数 $\phi(x)$ 映射到高维空间, 如果原始特征内积是 $<x_i,x_j>$, 映射后为 $<\phi(x_i),\phi(x_j)>$, 利用核函数 $K(x_i,x_j)=\phi(x_i)^T\phi(x_j)$ 代替原空间中的内积, 可以将原空间的数据隐含地表示在高维的特征空间中, 并且不需要知道具体的映射关系.

根据 MerCer 定理[4], 我们称一个核函数是有效的当且仅当对于训练样例 $\{x_1,x_2\cdots,x_m\}$, 其相应的核函数矩阵是对称半正定的.

在支持向量机中, 核函数及其参数的选择会对其产生很大的影响, 常用的核函数有以下四种:

1) 线性核函数

$$K(x_i,x_j)=x_i\cdot x_j \tag{1}$$

线性核函数就是原输入空间中任意两样本之间的内积, 也即原输入空间的恒等映射. 如果原输入空间中样本线性不可分, 通过线性核函数作用后仍无法达到线性可分的目的.

2) 多项式核函数

$$K(x_i,x_j)=\left[\gamma(x_i\cdot x_j)+1\right]^q,\ q\in Z^+ \tag{2}$$

多项式核函数具有良好的全局性, 即使相距很远的数据点都可以对核函数产生影响, 具有良好的泛化能力, 但是其局部性较差, 且多项式核函数的映射结果与 q 有很大关系, q 越大映射的维数越高, 计算量越大, q 过小会降低学习精度.

3) 高斯核函数

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{3}$$

高斯核函数又称径向基核函数, 是支持向量机中最常用的核函数, 具有良好的局部性, 只有相距很近的数据点才对核函数的值有影响, 学习能力较强, 但是其全局性较差, 不具有很好的泛化能力, 其泛化能力随着参数 σ 的增大而减弱.

4) Sigmoid 核函数

$$K(x_i,x_j)=\tanh(v(x_i\cdot x_j)+c) \tag{4}$$

Sigmoid 核函数来自于神经网络, 只有当参数 v 和 c 满足特定的条件时, Sigmoid 核函数才是半正定的, 所以在实际应用中一直受到限制.

2 混合核函数

2.1 多核学习方法

多核学习一般有以下 3 种方法[5,6]: 1)合成核方法: 通过多个核函数线性组合得到新的核函数; 2)多尺度多核学习方法[7]: 使用尺度核函数, 如高斯径向基核函数和小波核函数, 先用大尺度核拟合对应决策函数平滑区域的样本, 然后用小尺度核拟合决策函数变化的相对剧烈区域的样本; 3)无限核方法[8]: 由多个基本核函数的合法集合所构成的一个凸壳中找到某个核, 使其能最小化凸正则化函数, 与其他方法相比, 这个方法有一个独有的特征, 即上述基本核的个数可以是无限多个, 仅仅需要这些核是连续参数化的.

2.2 混合核函数

关于核函数的构造[9]主要依据以下定理:

定理 1. 设 K_1 和 K_2 都是核函数, 设常数 $\lambda\ge 0$, 则根据以下公式构造出来的函数均是核:

$$K(x_i,x_j)=K_1(x_i,x_j)+K_2(x_i,x_j)$$

$$K(x_i,x_j)=\lambda K_1(x_i,x_j)$$

$$K(x_i,x_j)=K_1(x_i,x_j)\cdot K_2(x_i,x_j)$$

核函数一般可以分为全局核函数和局部核函数: 全局核函数具有良好的全局性质, 如公式(2)中多项式核函数, 其图像如图 1 所示.

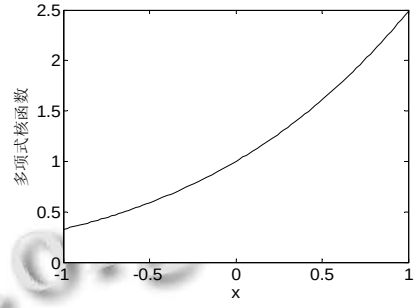

图 1　多项式核函数图

局部核函数具有良好的局部性质, 如公式(3)中高斯核函数, 其图像如图 2 所示.

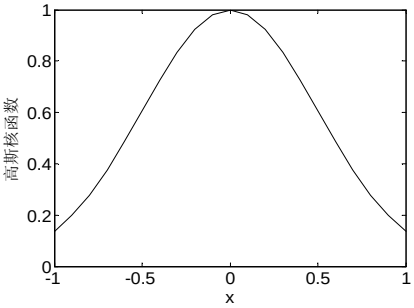

图 2　高斯核函数图

目前, 最简单也是最常用的混合核函数构造方法

就是核函数的线性组合形式, 由定理 1 可知线性组合取得的函数仍是核函数, 将全局核函数与局部核函数线性组合之后形成新的核函数, 不仅可以获得较好的学习能力, 也能得到较好的泛化能力[10], 本文将多项式核函数与高斯径向基核函数组合, 得到新的核函数:

$$K(x_i,x_j)=\lambda\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)+(1-\lambda)[\gamma(x_i\cdot x_j)+1]^q \quad (5)$$

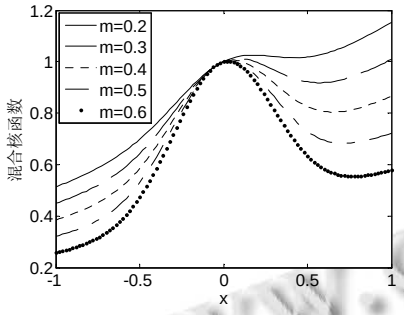

图 3 混合核函数图(图中 m = λ)

由图可知, 混合核函数结合了局部核函数和全局核函数的优点, 但是也引入了新的参数 λ, λ 的选择将直接影响各核函数的权重, 由图 3 可以看出, λ 的值得选取对预测结果会有较大影响.

3 现有的权重求解方法

目前对于权重参数的求解主要有穷举法和优化算法:

3.1 穷举法

该方法对 λ 取不同的值, 然后分别对样本集使用选取的 λ 进行 SVM 训练, 选择分类错误率最小的 λ 值作为结果. 典型的方法有交叉验证法[11].

利用交叉验证法来确定时, 首先需要给定一组的 λ 值, 然后分别对每一个 λ 建立 SVM 模型进行训练, 计算各自的实际风险估计性能指标, 选择性能指标最好的作为最终的混合核函数权重系数, 一般采用 k-折交叉验证法.

交叉验证实际上就是参数空间穷尽搜索法, 也就是说用枚举参数空间的每一组可能的参数去训练和测试 SVM, 找出效果最好的参数. SVM 的求解是比较耗时的, 当样本达到一定规模时, 交叉验证法将无法计算.

3.2 优化算法

对于参数的求解还有一种比较常用的方法是采用一些演化算法, 寻求最优解, 典型的有粒子群优化算法(Particle Swarm Optimization, PSO)[12], 这是一种基于群智能的随机优化算法, 每个粒子具有当前时刻自己的位置和速度, 通过迭代找到最优解, 每一次迭代中, 粒子通过跟踪两个“极值”来更新自己, 一个是粒子本身找到的最优解 pbest, 另一个是种群目前找到的最优解 gbest, 在找到这两个最优值时根据如下公式来更新自己的速度 V 和位置 X:

$$\begin{aligned}V_i^{t+1}&=V_i^t+c_1*rand()*(pbest_i^t-X_i^t)\\&\quad+c_2*rand()*(gbest_i^t-X_i^t)\\X_i^{t+1}&=X_i^t+V_i^t\end{aligned} \quad (6)$$

PSO 收敛速度较快, 但是在算法后期, 速度越来越小, 粒子群表现出强烈的趋同性, 易陷入局部极小点, 即具有“早熟”的缺陷, 虽然现在也出现了很多改进方法, 文献[13]和[14]中有介绍, 但是不能从根本上克服 PSO 算法的缺陷.

4 基于特征距离的权重求解方法

4.1 基于特征距离求权重方法

针对目前没有确定性方法确定混合核函数权系数的情况, 本文提出一种确定性算法求解权系数, 其原理如下:

核函数的本质就是某种映射关系的内积, 在混合核函数中, 经过映射之后的特征空间里, 样本 i, j 之间的距离可以定义为:

$$d=\|\phi(x_i)-\phi(x_j)\|^2 \quad (7)$$

在混合核函数中, 映射后的函数为 $\phi(x)=\lambda\phi_1(x)+(1-\lambda)\phi_2(x)$, 这里 ϕ_1 和 ϕ_2 分别表示使用高斯核函数和多项式核函数对应的函数映射, 分别用 $K_1(x_i,x_j)=\phi_1(x_i)\cdot\phi_1(x_j)$ 和 $K_2(x_i,x_j)=\phi_2(x_i)\cdot\phi_2(x_j)$ 表示高斯核函数和多项式核函数, 对(7)式展开后得到结果:

$$d(\lambda)=A\lambda^2+B\lambda+C \quad 0\le\lambda\le1 \quad (8)$$

其中

$$S=K_2(x_i,x_i)-2K_2(x_i,x_j)+K_2(x_j,x_j) \quad (9)$$

$$A=S-2K_1(x_i,x_j)+2 \quad (10)$$

$$B=-2S \quad (11)$$

$$C=S \quad (12)$$

在SVM的二分类问题中, 我们期望当两个样本属于同一类时其特征距离要尽量小, 而当样本属于不同类别时, 我们希望其特征距离尽量大, 即:

$$\begin{aligned}&\min d(\lambda), y_i y_j = 1\\&\min d(\lambda), y_i y_j = -1\end{aligned} \tag{13}$$

为了达到以上目的, 我们可以定义评估函数 $L(\lambda)$ 为异类样本间距与同类样本间距之差, 我们的目标是使其最大化, 即:

$$\begin{aligned}\max L(\lambda) &= \sum_{i=1}^{m}\sum_{j=1}^{i-1} d(\lambda)\Big|_{y_i y_j=-1} - \sum_{i=1}^{m}\sum_{j=1}^{i-1} d(\lambda)\Big|_{y_i y_j=1}\\&= -\sum_{i=1}^{m}\sum_{j=1}^{i-1} d(\lambda) y_i y_j\end{aligned} \tag{14}$$

将 $d(\lambda)$ 带入上式可以得到最终的优化函数是一个关于 λ 的一元二次多项式:

$$\max L(\lambda) = \sum_{i=1}^{m}\sum_{j=1}^{i-1} A y_i y_j \lambda^2 + \sum_{i=1}^{m}\sum_{j=1}^{i-1} B y_i y_j \lambda + \sum_{i=1}^{m}\sum_{j=1}^{i-1} C \tag{15}$$

要使 $L(\lambda)$ 最大化, 即求一元二次多项式的最大值问题, 其解为:

$$\lambda = -\frac{\sum_{i=1}^{m}\sum_{j=1}^{i-1} B y_i y_j}{2\sum_{i=1}^{m}\sum_{j=1}^{i-1} A y_i y_j} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{i-1} S y_i y_j}{\sum_{i=1}^{m}\sum_{j=1}^{i-1} (S - K_1(x_i, x_j) + 2) y_i y_j} \tag{16}$$

4.2 算法描述

本小节给出算法的伪码实现如下:

输入: 训练样本集{(x_i, y_i), i=1, ... ,m}

```
S=0;
G=0;
//求核函数及权重
for i=1 to m
    K1(i,i)=1;
    K2(i,i)=[gamma*(xi·xi)+1]^2 ;
    for j=1 to i-1
      //求高斯核函数
      K1(i,j)=exp(-1*||xi-xj||^2/2*sigma^2);
      K1(i,j)=K1(i,j);   //核矩阵对称
      //求多项式核函数
      K2(i,j)=[gamma*(xi·xj)+1]^2 ;
      K2(j,i)=K2(i,j);   //核矩阵对称
      //求 lammda 分子与分母
      S=S+[K2(i,i)+K2(j,j)-2K2(i,j)]*yi*yj;
      G=G+[2-K1(i,j)]*yi*yj;
    end for
end for
lammda=S/(S+G);
```

4.3 算法复杂度分析

从算法可以看出该方法求系数中只是用到了我们已经求得的核矩阵值, 不会增加额外的矩阵计算, 只是对已经得到的结果进行累加, 就算法本身而言, 其时间复杂度为 $O(m^2)$, m 为样本个数, 空间复杂度为 O(1), 整个计算过程中只需存储临时变量即可.

5 实验结果及分析

实验数据使用 UCI 数据集中 Musk Data Set, 该样本用于预测分子是否属于麝香类, 特征维数为 166, 用来描述分子的形状和构造, 样本标签为“1”和“-1”, 分别用来表示是否属于麝香类分子.

我们的目的是通过计算机建立模型, 对于每一个知道形状和构造的分子预测其是否属于麝香类. 该组样本中共有 476 组数据, 在实验中我们选取 300 个样本进行训练, 剩余 176 个样本用于预测. 为了便于实验, 先对样本进行归一化, 然后使用混合核函数进行训练和预测.

建立支持向量机模型, 对于支持向量机中参数 C 和 ξ 以及高斯核函数中 σ 使用交叉验证法确定, 多项式核函数中指数 q 取 2, 混合核函数中系数求解分别采用交叉验证法(5-folds)和标准 PSO 算法与本文的方法对比, 使用 MATLAB 工具进行实验.

使用交叉验证法计算时, λ 在 0-1 之间选取 20 个值, 然后取出训练时精度最高的 λ 作为预测使用的权重系数, 图 4 是交叉验证求解过程中 λ 变化与训练样本预测精度的关系图.

使用 PSO 算法计算时, 将预测精确度作为适应度函数, 选取初始群体个数为 10, 加速系数选为通用系数 2, 由于所求的参数只有一个, 所以迭代次数设为 10 即可, 图 5 是使用标准 PSO 算法求解后 λ 与迭代次数的关系图.

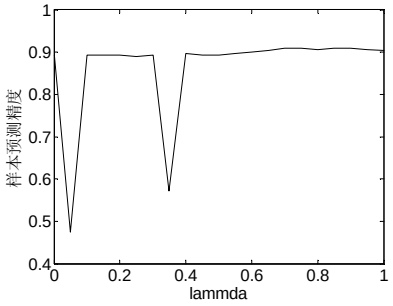

图 4　交叉验证法下 λ 与预测精度关系图

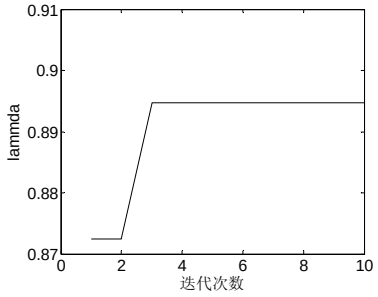

图 5 PSO 算法下 λ 与迭代次数关系图

从图中我们可以看出对于训练样本而言 λ 的最优值在 0.9 左右, 而使用本文提出的方法求得的 λ 结果与此一致, 且预测精度略高, 结果如表 1 所示:

表 1 不同方法求权重结果比较

使用方法	计算时间	预测精度	权重
交叉验证法	466.3088s	90.96%	0.9000
PSO 算法	521.2988s	90.30%	0.8947
基于距离法	165.2765s	90.97%	0.9595

从表 1 的结果可以看出, 我们的方法求得的结果较为准确, 在不降低预测精度的情况下大大减少了计算时间, 说明该方法具有一定的优势.

6 结语

基于支持向量机分类原理和核函数的知识, 本文提出一种基于特征距离的混合核函数权重求解方法, 通过实验表明该方法能够较快较准确的得到混合核函数的系数 λ, 与传统方法相比大大减少了计算时间.

在支持向量机模型中, 还有很多其他的参数, 这些参数的求解都没有确定的方法, 所以还需要进一步研究.

参考文献

1 Vapnik VN. Statistical Learning Theory. New York: Wiley, 1998.

2 Gonen M, Alpaydin E. Multiple kernel learning algorithms. Journal of Machine Learning Research, July 2011, 12(2): 2211–2268.

3 Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel based learning algorithms. IEEE Trans. on Neural Networks, 2001,12(2):181–201.

4 丁世飞,齐丙娟,谭红艳.支持向量机理论与算法研究综述.电子科技大学学报,2011,40(1):2–10.

5 汪洪桥,孙富春,蔡艳宁,陈宁,丁林阁.多核学习方法.自动化学报,2010,36(8):1037–1050.

6 介文博.基于多核学习的高性能核分类方法研究[硕士学位论文].上海:华东理工大学,2012.

7 Kingsbury N, Tay DBH, Palaniswami M. Multi-scale kernel methods for classication. Proc. of the IEEE Workshop on Machine Learning for Signal Processing. Washington D. C., USA. IEEE. 2005. 43–48.

8 Argyriou A, Hauser R, Micchelli CA, Pontil M. A DC-programming algorithm for kernel selection. Proc. of the 23rd International Conference on Machine Learning. Pittsburgh, USA: ACM, 2006. 41–48.

9 王国胜.核函数的性质及其构造方法.计算机科学,2006, 33(6):172–175.

10 Smits GF, Jordan EM. Improved SVM regression using mixtures of kernels. Proc. of the 2002 International Joint Conference on Neural Networks. Hawaii. IEEE. 2002. 2785–2790.

11 范永东.模型选择中交叉验证方法综述[硕士学位论文].太原:山西大学,2013.

12 Galewski MA. Modal parameters identification with particle swarm optimization. Progress in Mechanical Engineering and Technology. 2014. 597. 119–124.

13 Nguyen T, Li T, Zhang Z, Truong STK. A hybrid algorithm based on particle swarm and chemical reaction optimization. Expert Systems with Applications, April 2014, 41(5): 2134–2143.

14 基于 CPSO 的混合核函数 SVM 参数优化及应用.控制工程,2011,18(2):267–269.