文章编号:1000-5862(2013)06-0574-05

用核 K-means 聚类减样法优化半定规划支持向量机

何 慧[1],胡小红[2],覃 华[3],张 敏[3]

(1. 江西师范大学商学院电子商务系,江西 南昌 330022;2. 抚州市党校,江西 抚州 344000;
3. 广西大学计算机与电子信息学院,广西 南宁 530004)

摘要:提出了使用核空间 K-means 聚类算法从训练集中抽取特征边界支持向量集,在边界集上构造支持向量机的半定规划问题,由于边界集的规模比原始训练集要小,降低了半定规划支持向量机的规模,达到优化向量机的目的. 在 UCI 数据集上的实验结果表明:所提优化方法在求解多核半定规划向量机时,比原始方法获得几倍以上的速度提升,分类精度基本不变.

关键词:支持向量机;半定规划;核 K-means 聚类;减样

中图分类号:TP 309 **文献标志码**:A

0 引言

支持向量机(Support Vector Machines,SVM)是一种基于小样本统计学习的机器学习系统. 在使用 SVM 构造数据挖掘模型时,为 SVM 选择合适的核函数工作参数很重要,工作参数会通过核矩阵影响 SVM 模型的分类精度. 在选择 SVM 核函数工作参数问题上,较为常用的方法是经验值法和网格搜索(Grid Search)法[1-4],但这些方法缺乏成熟的理论作指导.

文献[5]提出了一种多核学习的支持向量机模型(Multiple Kernel SVM),把若干个不同工作参数的核矩阵通过最优线性组合后得到新的核矩阵,用新的核矩阵训练支持向量机,所得到的多核 SVM 模型的泛化能力优于普通的单核 SVM 模型,这种方法使用了最优组合理论来寻找多个核矩阵的最优组合系数,具有较好的实用性和数学理论支持. 对于给定的一组核矩阵,文献[5]通过半定规划方法来求解它们的最优线性组合系数,但当训练集较大时,所对应的半定规划问题是大规模的,求解时会花费较多的计算时间. 为解决此问题,文献[6]把多核 SVM 问题转化为 SILP 问题,用全局收敛算法求解 SILP,减少了计算时间,并且不影响多核 SVM 的精度.

本文根据支持向量机的工作原理,提出在核空间对训练集进行 K-means 聚类,用聚类算法从训练集中抽取对 SVM 分类超平面起决定性作用的边界支持向量,用小规模的边界支持向量构造出的支持向量机半定规划问题的规模也较小,求解此半定规划问题所花费的时间得到有效地缩减. 由于边界集上包含了对支持向量机决策起决定性作用的特征向量,所以新方法所获得的半定规划支持向量机的分类精度与原始训练集上的相当.

1 半定规划支持向量机

1.1 支持向量机模型

对于一个非线性可分的二分类问题,给定的训练数据集 X 记为 $X=\{(\boldsymbol{x}_i,y_i)\mid \boldsymbol{x}_i\in\mathbf{R}^d,y_i\in\{1,-1\},i=1,2,\cdots,n\}$,其中 $\boldsymbol{x}_i$ 为一个 d 维向量(输入样本). y_i 为样本 $\boldsymbol{x}_i$ 对应的类别标记,取值是 1 或 -1,表示正、负 2 种样本,n 为训练数据集的总样本数.

二分类支持向量机的训练学习原理是:把训练样本映射到高维 Hilbert 空间(核空间),在高维空间寻找一个分类超平面,使正、负训练样本点距离这个超平面尽可能的远. 支持向量机这种最大化样本间隔的思想可以转换为求解变量 b 的最优化问题[7-8]:

$$\begin{aligned}&\min\frac{1}{2}\|\boldsymbol{w}\|_2^2+\boldsymbol{C}\sum_{i=1}^{n}\xi_i,\\ &\text{SVM: } s.t.\quad y_i(\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i,\\ &\xi_i\geqslant 0,i=1,2,\cdots,n,\end{aligned}\qquad(1)$$

其中 w 为特征空间中分类超平面的系数(d 维向量),ξ_i 为考虑分类误差而引入的松驰因子,允许边界失败,b 为分类超平面的阈值(实数),$\varphi(\boldsymbol{x}_i)$ 为样

收稿日期:2013-07-27
基金项目:国家自然科学基金(61063032)和教育部人文社会科学研究规划基金(1YJAZH080)资助项目.
作者简介:何 慧(1978-),女,江西临川人,讲师,主要从事电子商务和数据挖掘的研究.

本x_i在高维特征空间的映射,可以是有限维,也可以是无穷维,C为一个与训练错误有关的惩罚因子.

由于(1)式中$\varphi(x_i)$为未知的非线性映射函数,很难直接求解.利用Lagrange乘子法将(1)式转化得到相应的对偶问题,具体如下[9]:

$$\begin{aligned} &\min \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}-\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha}, \\ &s.t.\quad \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0, \\ &\quad 0\leqslant\boldsymbol{\alpha}\leqslant\boldsymbol{C}, \end{aligned} \tag{2}$$

其中$\boldsymbol{\alpha}$为Lagrange乘子变,是原问题的变量,$\boldsymbol{y}$为一个n维的类别向量,取值是1或-1,$\boldsymbol{e}$为n维单位向量,$\boldsymbol{C}$为惩罚因子.$\boldsymbol{Q}$矩阵中的元素为

$$Q_{ij}=y_iy_j\boldsymbol{\varphi}^{\mathrm{T}}(x_i)\varphi(x_j), \tag{3}$$

$\boldsymbol{Q}$可以写为$\boldsymbol{Q}=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}$的形式,其中$\boldsymbol{B}=(y_1x_1,y_2x_2,\cdots,y_lx_l)_{n\times l}$是一个$n$行$l$列的矩阵,所以$\boldsymbol{Q}$是半正定的.

定义核函数$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$与内积$\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_i)\varphi(\boldsymbol{x}_j)$的关系为

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{\varphi}^{\mathrm{T}}(\boldsymbol{x}_i,\boldsymbol{x}_j)\varphi(\boldsymbol{x}_i,\boldsymbol{x}_j), \tag{4}$$

则可把(3)式写成核函数形式$Q_{ij}=y_iK(\boldsymbol{x}_i,\boldsymbol{x}_j)y_j$.

满足Mercer条件的核函数都应该是正半定的,例如RBF核函数、高斯核等,则此时(2)式是一个凸二次规划问题(Quadratic Programming,QP).

1.2　半定规划支持向量机模型

不妨定义一个核矩阵$\boldsymbol{K}$如下:

$$\boldsymbol{K}=\begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_n) \\ K(x_2,x_1) & K(x_2,x_2) & \cdots & K(x_2,x_n) \\ \cdots & \cdots & \cdots & \cdots \\ K(x_n,x_1) & K(x_n,x_2) & \cdots & K(x_n,x_n) \end{bmatrix}, \tag{5}$$

其中的$K(x_i,x_j)$用(4)式计算而得.不妨记:

$$\boldsymbol{Q}=G(\boldsymbol{K})=\mathrm{diag}(\boldsymbol{y})\boldsymbol{K}\mathrm{diag}(\boldsymbol{y}), \tag{6}$$

则(2)可以写成以下形式:

$$\begin{aligned} &\min \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}G(\boldsymbol{K})\boldsymbol{\alpha}-\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\alpha}, \\ &s.t.\quad \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\alpha}=0, \\ &\quad \boldsymbol{C}-\boldsymbol{\alpha}\geqslant 0, \\ &\quad \boldsymbol{\alpha}\geqslant 0. \end{aligned} \tag{7}$$

给定一组核矩阵,则多个核矩阵的线性组合可写为

$$\boldsymbol{K}=\sum_{s=1}^{m}\mu_s\boldsymbol{K}_s. \tag{8}$$

在(8)式中,对于一个给定的核函数工作参数,所对应的核矩阵记为$\boldsymbol{K}_s$,μ_s是这个核矩阵的线性组合系数.文献[9]证明了:若给定m个不同的核函数工作参数,则相应有m个不同的核矩阵$\boldsymbol{K}_s$,(8)式通过最优线性组合系数将这m个核矩阵相加,得到一个更优的核矩阵$\boldsymbol{K}$.将最优组合核矩阵$\boldsymbol{K}$代入(2)式中求解,即得到性能更优的SVM模型,称之为半定规划支持向量机模型.

通过半定规划来求解出多个核矩阵的最优线性组合系数μ_s.满足(7)~(8)式的半定规划问题可写为[10]:

$$\begin{aligned} &\min t \\ &\mathrm{s.t.}\begin{pmatrix} G(\boldsymbol{K})(\sum_{s=1}^{m}\mu_s\boldsymbol{K}_s) & (\boldsymbol{e}+\boldsymbol{v}-\boldsymbol{\sigma}+\lambda\boldsymbol{y}) \\ (\boldsymbol{e}+\boldsymbol{v}-\boldsymbol{\sigma}+\lambda\boldsymbol{y})^{\mathrm{T}} & t-2\boldsymbol{C}\boldsymbol{\sigma}^{\mathrm{T}}\boldsymbol{e} \end{pmatrix}\succeq 0 \\ &\quad \boldsymbol{e}^{\mathrm{T}}\boldsymbol{u}=1, \\ &\quad \boldsymbol{v}\geqslant 0, \\ &\quad \sigma\geqslant 0, \\ &\quad \mu\geqslant 0. \end{aligned} \tag{9}$$

求解(9)式,即可得到向量形式的多核最优线性组合系数μ,以及对偶变量$\boldsymbol{v}$、$\boldsymbol{\sigma}$、λ.根据最优组合系数求得最优组合核矩阵K及$G(K)$,由(7)式求得最优拉格朗日乘子$\boldsymbol{\alpha}^*$为[11]:

$$\boldsymbol{\alpha}^*=G(\boldsymbol{K})-1(\boldsymbol{e}+\boldsymbol{v}-\boldsymbol{\sigma}+\lambda\boldsymbol{y}). \tag{10}$$

从拉格朗日乘子$\boldsymbol{\alpha}^*$中任意选取一个正分量$0<\boldsymbol{\alpha}_j^*<\boldsymbol{C}$,由此计算出SVM分类阈值$\boldsymbol{b}^*$为

$$\boldsymbol{b}^*=\boldsymbol{y}_j-\sum_{i=1}^{n}y_i\boldsymbol{\alpha}_i^*K(\boldsymbol{x}_i,\boldsymbol{x}_j), \tag{11}$$

其中$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$是通过查最优组合核矩阵$\boldsymbol{K}$中的$(i,j)$元素而得.最后得到多核SVM的分类函数为

$$f(x)=\mathrm{sign}(\sum_{i=1}^{n}\boldsymbol{\alpha}_i^*y_iK(\boldsymbol{x}_i,\boldsymbol{x})+\boldsymbol{b}^*). \tag{12}$$

(9)式的约束中包含有核矩阵$\boldsymbol{K}$,当训练集较大时,核矩阵$\boldsymbol{K}$的维度很大,此时(9)式是一个大规模的半定规划问题,即使采用速度较快的原-对偶内点法来求解仍然会花费较长的时间.为减少求解(9)式的时间,本文引入核空间K-means聚类算法,从训练集中抽取对SVM分类超平面起决定性作用的边界支持向量作为训练集.由于边界支持向量一般只是训练集中的一部分,故用边界支持向量作为训练集所构造出的核矩阵$\boldsymbol{K}_s$比原始训练集上的$\boldsymbol{K}_s$规模要小,达到减小(9)式半定规划问题规模并减少求解时间的目的.由SVM的工作原理可知:对SVM的分类超平面起决定性作用的是边界支持向量,所以使用边界支持向量训练SVM,不会改变SVM的分类精度.

2　核空间K-means聚类算法

2.1　核距离

在核空间中,数据对象$\boldsymbol{x}=(x_1,x_2,\cdots,x_d)$和$\boldsymbol{y}=(y_1,y_2,\cdots,y_d)$之间的距离定义为:

$$d(x,y) = \|x-y\| = \sqrt{\sum_{i}^{d}(x_i-y_i)^2}. \quad (13)$$

在 SVM 中解决非线性的情况时,是将输入数据对象 X 映射到高维特征空间(可能是无穷维) H 中,即将 X 做变换 $\phi: \mathbf{R}^{(d)} \to H$,所有的样本都利用核函数映射到一个未知的特征空间里.数据对象(x,y)通过核函数映射到核空间后,它们在核空间中的核距离定义为:

$$d(x,y) = \|\varphi(x) - \varphi(y)\| = K(x,x) - 2K(x,y) + K(y,y)^{\frac{1}{2}}, \quad (14)$$

其中的 $K(\bullet,\bullet)$ 为(4)式中的核函数.

2.2 核空间 K-means 聚类算法

核空间 K-means 聚类算法参考迭代更新的聚类思想,通过核函数将数据集映射到核空间中,利用上述定义的核距离公式完成数据集在核空间的聚类.其算法描述如下(算法 I):

(a) 输入:包含 N 个数据样本的数据集 D、聚类数目 k,核函数 $K(x,y)$;

(b) 输出:使得目标函数最小的 k 个簇;

(c) 具体步聚力:

(ⅰ) 利用核函数 $K(x,y)$ 将数据对象 x_i 从原空间映射到高维核空间,映射后的样本对象为 $\varphi(x_i)$;

(ⅱ) 在核空间中随机选定 k 个对象作为初始聚类簇中心($mi\Phi,\cdots,mk\Phi$);

(ⅲ) 逐个将样本对象 $\varphi(xi)$ 按(3)式的核距离公式分配给距离最近的一个聚类核中心 $mi\Phi$;

(ⅳ) 重新计算聚类中心($mi\Phi,\cdots,mk\Phi$)和 Je 的值;

(ⅴ) 重复(ⅲ)和(ⅳ),直到 Je 值不变(或中心前后变化很小)为止;

(ⅵ) 输出聚类结果.

该聚类算法通过非线性映射能够较好地分辨、提取并放大样本有用的特征,从而实现更为准确的聚类,算法收敛速度也较快.

2.3 核空间边界支持向量的抽取

使用核空间聚类算法从大规模训练集中抽取边界支持向量的算法描述如下(算法 II):

设 M 为训练样本集,T_1 为训练集中 A 类样本集,T_2 为训练集中 B 类样本集合,$T_2 = M - T_1$,k 为聚类算法的参数,m 为控制所选样本数的参数,实验中的压缩比例即为 m/k.

(ⅰ) 对 T_1,T_2 分别利用算法 I 中的聚类算法进行聚类,并记录各簇的簇中心点.

(ⅱ) 利用核空间距离的公式计算 T_1 中各簇到 T_2 的距离 d_i,和 T_2 中各簇到 T_1 的距离 d_j,

$$d_i = \sum_{j=1}^{k} d(x_i,y_j) \; d_j = \sum_{i=1}^{k} d(y_j,x_i).$$

(ⅲ) 选取出 T_1 中距离 T_2 最近的前 m 个簇和 T_2 中距离 T_1 最近的前 m 个簇,并标记.

(ⅳ) 将 T_1、T_2 中有标记的样本取出作为支持向量机的边界支持向量集 M'.用 M' 对进行训练,得到 SVM 分类器.

2.4 采用核空间聚类的半定规划 SVM 学习算法

采用核空间聚类算法的半定规划 SVM 学习算法描述如下(算法 III):

(ⅰ) 调用算法 II 从原始训练集中通过聚类算法抽取边界支持向量集 M'.

(ⅱ) 给定一组核函数的工作参数,计算各个参数在训练集 M' 上的核矩阵 $\boldsymbol{K}_s$.

(ⅲ) 在训练集 M' 上构造(9)式中的半定规划问题.

(ⅳ) 用内点法求解半定规划问题(9),得到核矩阵最优组合系数,并由(8)式计算得到最优组合核矩阵 $\boldsymbol{K}$ 和 $G(\boldsymbol{K})$.

(ⅴ) 由最优组合核矩阵 $\boldsymbol{K}$,以及(9)式的求解结果,根据(11)式和(12)式得到多核支持向量机模型.

3 实验

为了验证核空间聚类的半定规划 SVM 学习算法的有效性,实验软硬件环境如下:实验测试软硬件条件是 Pentium dual-core,2.7 GHz CPU,2GB 内存的PC机,windows XP系统,Java语言集成开发环境,采用 Matlab 6.5 实现算法及支持向量机模型,选用 SDPA [10] 求解半定规划问题.选用的 RBF 核函数为

$$K(x,y) = \exp(-\gamma\|x-y\|^2). \quad (15)$$

RBF-SVM 的主要工作参数有:C 是公式(1)中的惩罚因子,γ 是公式(15)中的核函数参数.RBF核函数的参数 γ 对 RBF-SVM 算法的分类精度影响较大,所以实验固定参数 $C=1$,变化 RBF 核函数参数 γ,产生一组核矩阵 K_s,通过实验检测本文算法查找核矩阵最优线性组合参数的时间消耗及多核 SVM 分类精度的变化.

选取 3 个 UCI 数据集 Breast Cancer(BRC)、Ionosphere(IOS)、Diabetes(DIB),以及 KDD CUP 1999 中的 DOS 网络入侵检测数据集(DOS)来测试本文算法的可行性.

利用本文核空间聚类算法约简训练集后的结果见表1.其中,“训练样本数”表示是从原始数据集中选出的训练样本数目.“约减后样本”表示采用本文核空间聚类算法从训练样本中抽取到的边界支持向量样本数.“约简比(%)”的计算方法是:约减后样本数/训练样本数×100%,表示约简后的样本占原

训练样本的百分率. 从表 1 中可看出,约减后得到的边界支持向量样本数比训练集样本有大幅减少,当训练集样本规模较大时,约简效果更加明显,有利于大幅减少核矩阵的维度.

表1　训练样本的约减

数据集	训练样本数	约减后样本	约简比 /%
BRC	387	37	9.50
IOS	284	62	21.80
DIB	515	105	20.40
DOS	64 571	278	0.43

求解多核最优组合系数的结果见表2. 其中,"3 个参数值 γ" 这一列是为数据集选取的 3 个 RBF 核工作参数 γ,根据这 3 个参数,由(15) 式和(5) 式可得到训练集上 3 个不同的核矩阵. "半定规划求解时间" 这一列中有 3 个子列,其中 "前" 这一列中的数据是使用表 1 中的 "训练样本数" 中的训练样本构造 3 个核矩阵,求解相应的半定规划问题(9) 所用的时间,单位为 s. "后" 这一子列中的数据是使用表 1 中 "约简后样本" 中的训练样本构造 3 个核矩阵,求解相应的半定规划问题(9) 所用的时间,单位为 s. "加速比" 这一子列中的数据是用 "前" 列中的数据除以 "后" 列中的数据所得,用于描述用本文核空间聚类算法约简数据集后,求解多核组合半定问题时所花费的时间减少的程度. 从加速比数据上看,用核空间 K-means 聚类算法约简训练集后,求解多核组合系数的时间成倍地减少,例如 DOS 训练集约简前求多核组合系数花费近 9 min,约简后仅用近 24 s,求解速度提升了近 23 倍,这对处理大规模多核支持向量机问题是有益的. 从表 1 的 "约简比" 列和表 2 的 "加速比" 列中的数据上看,约简比越小,表示边界支持向量占训练样本数据比率越小,则核矩阵的规模就会大幅减小,求解多核组合系数时速度提升更明显.

表2　多核组合系数的求解时间

数据集	3 个参数值 γ			半定规划求解时间 /s		
				前	后	加速比
BRC	0.010 0	0.111	1	403.3	13.5	29.9
IOS	0.029 4	0.100	1	27.9	4.4	6.3
DIB	0.010 0	0.125	1	461.4	65.9	7.0
DOS	0.032 3	0.100	1	536.6	23.7	22.6

用约简后的训练样本构造核矩阵时,所求得的多核最优组合系数以及 SVM 模型的分类精度见表 3. "最优组合系数 μ" 这一列中的数据是用半定规划法求得的各数据集 3 个核矩阵的最优组合系数. 如果一个数据集中的某个组合系数太小,表示这个系数对应的核函数工作参数 γ 不适合此数据集,应该另选一个工作参数重新组合[12],例如 DIB 数据集的第 2 个组合系数为 "0.000 95",这个系数表示相应的核矩阵所占的权重很小,在表 2 中相应的 γ = 0.125 这个参数不适合此数据集,应该另选一个 γ 参数重新进行组合计算.

表3　多核最优组合系数及多核 SVM 分类精度

数据集	最优组合系数 μ			分类精度 /%	
				前	后
BRC	0.004 95	0.008 34	0.986 71	96.32	96.52
IOS	0.414 82	0.288 17	0.297 01	88.20	88.30
DIB	0.879 29	0.000 95	0.119 76	70.80	70.80
DOS	0.650 31	0.140 12	0.209 56	98.50	98.70

在表 3 中, "分类精度" 这一列中包含 "前"、"后" 2 个子列,分别表示训练集约简前、后,所得多核 SVM 模型在测试集上的分类精度. 并对改进算法复杂度与原有算法复杂度进行对比分析. 测试集样本也是从原始数据集中抽取,与训练集样本不重复,以便更客观地考察多核 SVM 模型的泛化能力. 把 IOS 数据集通过复制的方法,增加其数据量,分别得到 2 倍、4 倍、8 倍、16 倍的数据量的数据集,各记为 IOS-2、IOS-4、IOS-8、IOS-16,用改进前和改进后的算法对这些数据集聚类计算,所需要的计算时间如表 4 所示.

表4　算法改进前后的计算时间比较　　单位: s

	IOS	IOS-2	IOS-4	IOS-8	IOS-16
改进前算法	27.9	34.5	43.2	61.6	95.2
改进后算法	4.4	7.2	11.3	15.7	21.5

表 4 可用图 1 表示.

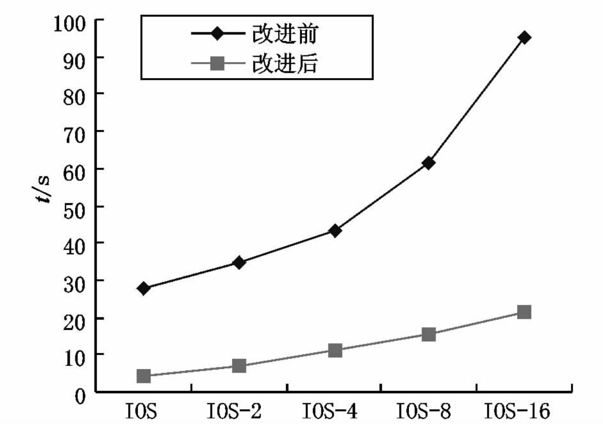

图1　不同规模算例的计算时间

由图 1 可知,当 IOS 数据集的规模呈 $2N$ 指数级增长时,改进前的算法计算时间呈非线性增长趋势,而改进后的算法则呈近似线性的增长趋势,所以改进后的本文算法具有近似线性的计算时间复杂度.

从分类精度数据上看,本文使用核空间 K-means 聚类算法约简训练集后,所获得的多核 SVM 模型的分类精度基本上没有降低,因为本文算法没有丢失 SVM 最关键的边界支持向量. 从算法复杂度上看,改进后算法其训练时间比改进前算法的训练

时间有大幅减少,当训练集样本规模越大,训练时间减少的幅度更加明显.

4 总结

在半定规划支持向量机模型中,求解速度会受到训练集规模的影响,当训练集规模较大时,核矩阵的维数较高,导致相应的半定规模问题属于大规模问题,求解难度高,时间花费大.本文根据 SVM 的工作原理,提出使用核空间 K-means 聚类算法约简训练集,降低训练集和核矩阵的规模,进而降低了 SVM 半定规划问题的规模.实验结果表明:经本文算法优化后,在求解多核半定规划问题时可获得显著的速度提升.由于约简后所得的边界支持向量不改变 SVM 的分类超平面,所以用约简后的训练集所求得的半定规划 SVM 模型分类精度基本不变,本文使用核空间 K-means 聚类优化大规模多核 SVM 的思路是可行的,且聚类计算时间也大大的减少.

5 参考文献

[1] Bao Yukun,Liu Zhitao. A fast grid search method in support vector regression forecasting time series [C]. Berlin: Springer Berlin Heidelberg,2006:504-511.

[2] Lazar A. Income prediction via support vector machine [EB/OL]. [2012-10-18]. http://dblp.uni-trier.de/rec/bibtex/conf/icmla/Lazar04.

[3] Chih-Jen Lin. Binary-class cross validation with different criteria [EB/OL]. [2010-02-04. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/eval/index.html.

[4] 郑成勇.小波与多小波变换在支持向量机人脸识别中的研究 [J].江西师范大学学报:自然科学版,2008,32(1):77-80.

[5] Gert Lanckriet,Nello Cristianini,Peter Bartlett,et al. Learning the Kernel matrix with semidefinite programming [J]. Journal of Machine Learnig Research,2004(5):27-72.

[6] Li Kan. A new algorithm for solving multiple Kernel problem as SLIP [C]. Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control,2008:409-415.

[7] Downs T,Gates T E,Masters A. Exact simplification of support vector solutions [J]. Journal of Machine Learning Research,2001(2):293-297.

[8] Haasdonk B. Feature space interpretation of SVMs with indefinite kernels [J]. Pattern Analysis and Machine Intelligence,IEEE Transactions,2005,27(4):482-492.

[9] Jair Cervantes,Li Xiaoou,Yu Wen,et al. Support vector machine classification for large data sets via minimum enclosing ball clustering [J]. Neurocomputing,2008(416),71:611-619.

[10] Nello Cristianini,Jaz Kandola,Andre Elisseeff,et al. On Kernel target alignment [J]. Innovations in Machine Learning,2006,194:205-256.

[11] Yang Shuzhong,Luo Siwei. Learning SVM Kernel with semi-definite programming [J]. Advances in Natural Computation,2005,3601:710-715.

[12] Makoto Yamashita,Katsuki Fujisawa,Masakazu Kojima. Implementation and evaluation of SDPA 6.0(semi-definite programming algorithm 6.0) [J]. Optimization Methods and Software,2003,18(4):491-505.

Using Kernel K-Means Clustering Reducing Method for the Optimization of Semi-Definite Programming SVM

HE Hui[1],HU Xiao-hong[2],QIN Hua[3],ZHANG Min[3]

(1. Electronic Commerce Department of Business School,Jiangxi Normal University,Nanchang Jiangxi 330022,China;

2. CPC Fuzhal Municipal Party,Fuzhou Jiangxi 344000,China;

3. Department of Computer and Electronics Information,Guangxi University,Nanning Guangxi 530004,China)

Abstract: Kernel K-means clustering method is proposed for abstracting the border support vector data set from training data set. The semi-definite programming SVM is solved on border set. The SVM scale is reduced as the border set is less than the original training data set,and the optimization of semi-definite programming is implemented. The experimental results on UCI data set show that the new SVM training time is several times less than the original one and the classification accuracy of new SVM is equals to original one.

Key words: SVM; semidefinite programming; Kernel K-means clustering; reducing

(责任编辑:冉小晓)