ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 6,Nov. 2018,pp. 1068－1076
DOI:10. 16337/j. 1004-9037. 2018. 06. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

并行效率敏感的大规模 SVM 数据分块数选择

张　闯　廖士中

(天津大学计算机科学与技术学院,天津,300350)

摘　要:数据分块数的选择是并行/分布式机器学习模型选择的基本问题之一,直接影响着机器学习算法的泛化性和运行效率。现有并行/分布式机器学习方法往往根据经验或处理器个数来选择数据分块数,没有明确的数据分块数选择准则。提出一个并行效率敏感的并行/分布式机器学习数据分块数选择准则,该准则可在保证并行/分布式机器学习模型测试精度的情况下,提高计算效率。首先推导并行/分布式机器学习模型的泛化误差与分块数目的关系。然后以此为基础,提出折衷泛化性与并行效率的数据分块数选择准则。最后,在 ADMM 框架下随机傅里叶特征空间中,给出采用该数据分块数选择准则的大规模支持向量机实现方案,并在高性能计算集群和大规模标准数据集上对所提出的数据分块数选择准则的有效性进行实验验证。

关键词:大规模支持向量机;模型选择;数据分块;交替方向乘子法;随机傅里叶特征

中图分类号:TP181　　**文献标志码**:A

Parallel Efficiency Sensitive Criterion of Data Segmentation for Large-Scale SVM

Zhang Chuang, Liao Shizhong

(School of Computer Science and Technology, Tianjin University, Tianjin, 300350, China)

Abstract: Data segmentation is one of critical issues of model selection of parallel/distributed machine learning, which has impacts on generalization performance and parallel efficiency of parallel/distributed machine learning. Existing approaches to data segmentation of parallel/distributed machine learning are dependent on empirical evidences or on the number of the processors without explicit criterion. In this paper, we propose a parallel efficiency sensitive criterion of data segmentation with generalization theory guarantee, which improves the computational efficiency of parallel/distributed machine learning while retaining test accuracy. We first derive a generalization error upper bound with respect to the block number of the data segmentation. Then we present a data segmentation criterion that is a trade-off between the generalization error and the parallel efficiency. Finally, we implement large-scale Gaussian kernel support vector machines (SVMs) in the random Fourier feature space with the alternating direction method of multipliers (ADMM) framework on high-performance computing clusters, which adopt the proposed data segmentation criterion. Experimental results on several large-scale benchmark datasets show that the proposed data segmentation criterion is effective and efficient for the large-scale SVMs.

Key words: large-scale support vector machines; model selection; data segmentation; alternating direction method of multipliers; random Fourier features

基金项目:国家自然科学基金(61673293)资助项目。
收稿日期:2017-06-09;**修订日期**:2017-07-09

引　　言

支持向量机(Support vector machines,SVMs)是基本机器学习模型和有效的数据挖掘方法之一。核 SVM 通过引入核技巧实现应用线性方法来学习非线性关系的途径,但 SVM 的求解是一个二次规划问题,时间复杂度为 $O(n^3)$,空间复杂度为 $O(n^2)$,其中 n 为训练集规模[1-2],这成为发展大规模 SVM 的主要瓶颈。

为发展大规模 SVM,文献[3]提出基于核矩阵近似的并行 SVM 算法,引入核缓存策略来并行计算核矩阵,数据分块数设为 l/d,其中,l 为训练集规模,d 为核缓存区大小。近年来,应用随机特征映射近似核方法的研究引起了广泛关注。Rahimi 等人提出利用随机傅里叶特征映射近似高斯核函数,进而在显式随机特征空间中应用线性 SVM 来一致逼近核诱导特征空间中的高斯核 SVM[4-5],成为提升 SVM 可扩展性的有效方法。Feng 等人提出一种新的随机特征映射方法[6],利用有符号循环随机矩阵代替无结构随机矩阵来投影数据。该方向的工作进一步发展了有效的近似方法,时间复杂度可降至与数据规模呈对数线性,同时也发展了大规模核方法[7]。文献[8]通过引入半宽因子,构造高斯区间核 SVM 模型,发展了针对区间型数据的高效分类方法。

交替方向乘子法(Alternating direction method of multipliers, ADMM)提供一个简单且强大的并行/分布式计算框架,可将大规模问题分解为多个小规模子问题,进而相互协调且并行一致地求解原问题[9]。在该框架下,文献[10]提出基于并行/分布式 ADMM 的线性 SVM 算法。该算法中的数据分块数依据经验来选择,没有探讨数据分块数对模型泛化性和计算效率的影响。文献[11]将随机特征映射与并行/分布式 ADMM 相结合,提出基于核方法统计模型的大规模训练框架。该框架适合多种统计学习任务,分块数目经验地取为 D/d,其中,D 为随机特征维度,d 为训练数据维度。矩阵分解与填充广泛应用于推荐系统中,分布式矩阵分解得到了广泛关注。Yu 等人[12]研究随机 ADMM 框架下分布式求解矩阵分解问题,矩阵划分块数与集群节点个数相同,没有讨论分块数目对均方根误差的影响。Zhang 等人[13]提出基于分解法的可扩展核岭回归算法,通过适当的参数调节,只要分块数目不是太多,该方法可以提高计算速度,保持统计最优性,进而得到有效的一致模型估计量。

已有的并行/分布式机器学习方法中,数据分块工作没有明确的分块数选择准则,也缺乏基本的理论保证。针对这一问题,提出大规模并行效率敏感的数据分块数选择准则。该准则以并行/分布式机器学习的泛化误差与数据分块数之间的关系为基础,折衷泛化误差与并行效率,可在保证并行/分布式机器学习测试精度的条件下,提高计算效率。在 ADMM 框架下的随机傅里叶特征空间中,采用所提出的数据分块数选择准则实现一个大规模支持向量机模型。实验结果表明,该准则在保证大规模支持向量机测试精度的同时,仍可进一步提高计算效率。

1　相关工作

1.1　交替方向乘子法

当目标函数满足可加性时,并行/分布式 ADMM 等式约束问题[9]为

$$\min_{\boldsymbol{\omega},\boldsymbol{z}} \sum_{i=1}^{B} f_i(\boldsymbol{A}_i\boldsymbol{\omega}_i - \boldsymbol{b}_i) + g(\boldsymbol{v}), \text{s.t.}\ \boldsymbol{\omega}_i - \boldsymbol{v} = 0, \quad i = 1,\cdots,B$$

其中:$\boldsymbol{\omega}_i \in \mathbf{R}^d$,$\boldsymbol{v} \in \mathbf{R}^d$,$\boldsymbol{A}_i \in \mathbf{R}^{n_i \times d}$,$\boldsymbol{b}_i \in \mathbf{R}^{n_i}$ 且 $\sum_{i=1}^{B} n_i = N$,N 为数据规模。缩放形式的增广拉格朗日函数为

$$\mathcal{L}_i(\boldsymbol{\omega}_i,\boldsymbol{v},\boldsymbol{u}_i) = f_i(\boldsymbol{A}_i\boldsymbol{\omega}_i - \boldsymbol{b}_i) + g(\boldsymbol{v}) + \frac{\rho}{2}\|\boldsymbol{\omega}_i - \boldsymbol{v} + \boldsymbol{u}_i\|_2^2 - \frac{\rho}{2}\|\boldsymbol{u}_i\|_2^2$$

其中：$\rho>0$ 为惩罚系数，$\boldsymbol{u}_i=\boldsymbol{y}/\rho$ 为缩放对偶变量，$\boldsymbol{y}$ 为拉格朗日乘子向量。每个从进程并行更新局部变量 $\boldsymbol{\omega}_i,\boldsymbol{u}_i$。主进程汇聚 $\boldsymbol{\omega}_i,\boldsymbol{u}_i$，更新全局变量 $\boldsymbol{v}$，并广播给从进程。并行/分布式 ADMM 交替优化过程见算法 1。

算法 1 分布式交替方向乘子法(D-ADMM)

输入：函数 f,g，矩阵 $\boldsymbol{A}$，分块数备选集 B，惩罚系数 $\rho>0$。

(1) 初始化：$\boldsymbol{\omega}^0,\boldsymbol{v}^0,\boldsymbol{y}^0$；

(2) repeat

(3) $\boldsymbol{\omega}_i^{k+1}:=\underset{\boldsymbol{\omega}_i}{\operatorname{argmin}} f_i(\boldsymbol{A}_i\boldsymbol{\omega}_i-\boldsymbol{b}_i)+\frac{\rho}{2}\left\|\boldsymbol{\omega}_i-\boldsymbol{v}+\boldsymbol{u}_i^k\right\|_2^2$；

(4) $\boldsymbol{v}^{k+1}:=\underset{\boldsymbol{v}}{\operatorname{argmin}} g(\boldsymbol{v})+\frac{B\rho}{2}\left\|\boldsymbol{v}-\bar{\boldsymbol{\omega}}^{k+1}-\bar{\boldsymbol{u}}^k\right\|_2^2$；

(5) $\boldsymbol{u}_i^{k+1}:=\boldsymbol{u}_i^k+\boldsymbol{\omega}_i^{k+1}-\boldsymbol{v}^{k+1}$；

(6) until 满足终止条件。

输出：$\boldsymbol{\omega}^*,\boldsymbol{v}^*,\boldsymbol{y}^*$。

D-ADMM 的终止条件为

$$\text{原残差}\left(\sum_{i=1}^{B}\left\|\boldsymbol{\omega}_i^{k+1}-\boldsymbol{v}^{k+1}\right\|_2^2\right)^{1/2}\leqslant\varepsilon^{\text{pri}}\ \text{且对偶残差}\ \rho\sqrt{N}\left\|\boldsymbol{v}^{k+1}-\boldsymbol{v}^k\right\|_2\leqslant\varepsilon^{\text{dual}}$$

其中：$\varepsilon^{\text{pri}}>0,\varepsilon^{\text{dual}}>0$ 分别为原问题和对偶可行性条件的误差，定义为

$$\varepsilon^{\text{pri}}=\sqrt{N}\varepsilon^{\text{abs}}+\varepsilon^{\text{rel}}\max\{\left\|\bar{\boldsymbol{\omega}}^{k+1}\right\|_2,\left\|\boldsymbol{v}\right\|_2\},\varepsilon^{\text{dual}}=\sqrt{N}\varepsilon^{\text{abs}}+\varepsilon^{\text{rel}}\rho\left\|\bar{\boldsymbol{u}}^{k+1}\right\|_2$$

其中：$\varepsilon^{\text{abs}}>0,\varepsilon^{\text{rel}}>0$ 分别为绝对误差和相对误差[9]。

满足如下两个假设的条件时，D-ADMM 收敛[9,14]。

假设 1 扩展实值函数 $f:\mathbf{R}^n\to\mathbf{R}\cup\{+\infty\},g:\mathbf{R}^m\to\mathbf{R}\cup\{+\infty\}$ 是封闭、适定且凸的[15]。

假设 2 增广拉格朗日函数 $\mathscr{L}$ 存在鞍点。

1.2 随机傅里叶特征映射

高斯核函数是一种通用的平移不变核，定义为

$$k(\boldsymbol{x},\boldsymbol{y})=\exp(-\gamma\left\|\boldsymbol{x}-\boldsymbol{y}\right\|_2^2)\tag{1}$$

式中 $\gamma=1/2\sigma^2$。

可通过高斯核函数的傅里叶逆变换得到 $\boldsymbol{\omega}$ 的概率密度函数 $p(\boldsymbol{\omega})$，$\boldsymbol{\omega}\sim$ 正态分布 $\mathcal{N}(0,2\gamma\boldsymbol{I})$，其中 $\boldsymbol{I}$ 为单位矩阵。易知

$$k(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{E}\left[\mathrm{e}^{-\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y})}\right]=\boldsymbol{E}_{\omega,b}\left[\sqrt{2}\cos(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b)\sqrt{2}\cos(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{y}+b)\right]$$

令 $T_{\omega,b}(\boldsymbol{x})=\sqrt{2}\cos(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}+b)$ 得 $k(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{E}\left[\langle T_{\omega,b}(\boldsymbol{x}),T_{\omega,b}(\boldsymbol{y})\rangle\right]$。

可见，$\langle T_{\omega,b}(\boldsymbol{x}),T_{\omega,b}(\boldsymbol{y})\rangle$ 是高斯核函数 (1) 的无偏估计。通过标准蒙特卡洛积分近似高斯核，构造如下随机傅里叶特征映射[4-5]

$$\Phi:x\to\sqrt{\frac{2}{D}}\cos(\boldsymbol{T}\boldsymbol{x}+\boldsymbol{b})\tag{2}$$

式中：D 为随机特征维度，高斯随机矩阵 $\boldsymbol{T}\in\mathbf{R}^{D\times d}$，$\boldsymbol{T}_i\sim\mathcal{N}(0,2\gamma\boldsymbol{I})$，$\boldsymbol{b}$ 为随机向量，$\boldsymbol{b}_i\sim$ 均匀分布 $U(-\pi,\pi),i=1,\cdots,D$。则 $k(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{E}\left[\langle\Phi(\boldsymbol{x}),\Phi(\boldsymbol{y})\rangle\right]$。

2 并行/分布式机器学习模型泛化误差分析

本节推导并行/分布式机器学习模型数据分块数与泛化误差之间的关系。

不失一般性，界定以下假设条件：

假设 3　$f^* \in C(X)$ 且 $\|f^*\|_\infty \leqslant M$。其中,$C(X)$ 是 X 上的连续函数空间,$\|\cdot\|_\infty$ 为上确界范数。

假设 4　损失函数 $l(\cdot)$ 为非负 L-Lipschiz 连续的凸函数。$\mathcal{H}_K$ 是一再生核 Hilbert 空间 (RKHS)。对任意 $f_1, f_2 \in \mathcal{H}_K$,存在常数 $L>0$,使得

$$|l(f_1,z)-l(f_2,z)| \leqslant L\|f_1-f_2\|_\infty$$

假设 5　任意 $g \in C(X)$,$\varepsilon>0$,存在 $f \in \mathcal{H}_K$,使得 $\|f-g\|_\infty < \varepsilon$。令 $B_R = \{f \in \mathcal{H}_K, \|f\|_\infty \leqslant R\}$,$R>0$。存在常数 $C_0, s>0$,使得

$$\mathcal{N}_\infty(F,r) \leqslant \exp(C_0 r^{-s})$$

其中:$\mathcal{N}_\infty(F,r)$ 表示集合 F 半径为 r 球的覆盖数。

下面分析并行/分布式机器学习模型的泛化误差。可将泛化误差分解为采样误差、假设误差和近似误差 3 部分,则有

$$\varepsilon(f_B)-\varepsilon(f^*) = \underbrace{\varepsilon_S(f)-\varepsilon(f)+\varepsilon(f_B)-\varepsilon_S(f_B)}_{\text{采样误差}} + \underbrace{\varepsilon_S(f_B)-\varepsilon_S(f)}_{\text{假设误差}} + \underbrace{\varepsilon(f)-\varepsilon(f^*)}_{\text{近似误差}}$$

其中:$f^* \in C(X)$,$f \in \mathcal{H}_K$ 是在样本 S 上学习的结果,$f_B \in \mathcal{H}_K$ 是把样本分成 B 块后学习的结果,$\varepsilon(\cdot)$ 为期望误差,$\varepsilon_S(\cdot)$ 为经验误差。令样本规模为 N,基于以上假设和分析可给出如下泛化误差分析结果。

引理 1[16]　假设 3—5 成立,$M'=\max\{2M, \|f-f^*\|_\infty\}$。对任意 $0<\delta<1$,有

$$\Pr\left\{\varepsilon_S(f)-\varepsilon(f)+\varepsilon(f_B)-\varepsilon_S(f_B) \leqslant 6M'L\left(\frac{G_1(N/B,\delta)}{N/B}+\frac{G_2(N/B,\delta)}{\sqrt{N/B}}\right)+\frac{1}{(N/B)^\tau}\right\} \geqslant 1-\delta \quad (3)$$

式中:$\tau=1/(s+2)$;$P(N,\delta)=C_0(8LMN^\tau)^s-\log\delta$;$G_1(N,\delta)=P(N,\delta/2)+\log(2/\delta)$;$G_2(N,\delta)=\sqrt{P(N,\delta/2)}+\sqrt{\log(2/\delta)}$。

引理 2[16]　假设 3—5 成立,对任意 $0<\delta<1$,有

$$\Pr\left\{\varepsilon_S(f_B)-\varepsilon_S(f) \leqslant 6LM'\left(\frac{G_1(N/B,\delta/2)}{N/B}+\frac{G_2(N/B,\delta/2)}{\sqrt{N/B}}\right)+\frac{1}{(N/B)^\tau}+2\lambda\|f\|_2^2\right\} \geqslant 1-\delta \quad (4)$$

式中 M',G_1 和 G_2 定义同引理 1。

定理 1　假设 3—5 成立,当 N 足够大时,对任意 $0<\delta<1$,$f \in \mathcal{H}_K$,有

$$\Pr\left\{\varepsilon(f_B)-\varepsilon(f^*) \leqslant 24LM\left(\frac{G_1(N/B,\delta/4)}{N/B}+\frac{G_2(N/B,\delta/4)}{\sqrt{N/B}}\right)+\frac{2+L}{(N/B)^\tau}+2\lambda\|f\|_2^2\right\} \geqslant 1-\delta \quad (5)$$

式中:$\|f-f^*\|_\infty \leqslant 1/N$,$G_1$ 和 G_2 定义同引理 1。

证明　由假设 3 和假设 5 成立可知,对任意 $N \geqslant 1$,存在 $f \in \mathcal{H}_K$,有

$$\|f-f^*\|_\infty < 1/N$$

由假设 4 成立,损失函数 $l(\cdot)$ 是 L-Lipschiz 连续的,有

$$\varepsilon(f)-\varepsilon(f^*)=E_z[l(f,z)-l(f^*,z)] \leqslant L\|f-f^*\|_\infty < L/N \leqslant L/(N/B)^\tau \quad (6)$$

所以,近似误差上界为 $L/(N/B)^\tau$。当 N 足够大时,$\|f-f^*\|_\infty < 1/N \to 0$。那么

$$M'=\max(2M, \|f-f^*\|_\infty)=2M \quad (7)$$

将式(7)分别代入采样误差式(3)和假设误差式(4)并与近似误差式(6)求和,可得式(5)。

定理 1 表明,分块模型泛化误差界为 $O((N/B)^{-\tau})$。分块数 B 越少,每个进程上数据规模 N/B 越大,分块模型泛化误差越小,但每个进程的计算时间越长。所以,选择最优分块数 $\hat{B}$,可在保证并行/分布式机器学习模型泛化性的同时提高计算效率。

3　数据分块数选择准则

给定训练数据 $A=\{r_i=(x_i,y_i), i=1,2,\cdots,N\}$,其中,$x_i \in \mathbf{R}^d$,$d$ 为训练数据维度,N 为训练集规模。

由经验风险最小化[17]可得

$$\boldsymbol{f}=\underset{f\in\mathcal{H}_K}{\arg\min}\frac{1}{N}\sum_{i=1}^{N}l(f,\boldsymbol{r}_i)$$

其中：f 为再生核希尔伯特空间 $\mathcal{H}_K$ 中的任意假设，$l(\cdot)$ 为非负凸损失函数。

对训练数据 $\boldsymbol{A}$ 按行划分为 B 块[18]，每个子数据块 $\boldsymbol{A}_i\in\mathbf{R}^{|B_i|\times d}$，其中，$|B_i|$ 为子数据块规模，且 $\sum_{i=1}^{B}|B_i|=N$。

定义分块经验风险最小化，可得

$$f_B=\underset{f\in\mathcal{H}_K}{\arg\min}\sum_{i=1}^{B}l_i(f,\boldsymbol{A}_i)$$

其中 $l_i(\cdot)$ 为第 i 个子数据块的平均经验风险。

并行算法效率[19]定义为

$$E=\frac{T_s}{T_pB}$$

其中：T_s 为串行时间，T_p 为并行时间，B 为进程数。

为了权衡模型的泛化性和并行效率，提出并行效率敏感的数据分块数选择准则

$$\hat{B}=\underset{B\in\mathcal{B}}{\arg\min}\sum_{i=1}^{B}l_i(f_B,\boldsymbol{A}_i)+\eta E(B)$$

其中：$\mathcal{B}$ 为分块数备选集，η 为惩罚系数，$\delta\leqslant E(B)<1$，δ 为并行效率下界。

4 大规模支持向量机

本节采用所提出的数据分块数选择准则来构造一个大规模支持向量机模型。

在并行/分布式 ADMM 框架下的随机傅里叶特征空间中实现大规模支持向量机模型。给定标签数据集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}\in(X\times Y)^N$，其中，$X$ 表示输入域，Y 为输出域，$x_i\in\mathbf{R}^d$，标签 $y_i\in\{-1,+1\}$，N 为训练集规模。

由随机傅里叶特征映射式(2)，得到随机特征映射矩阵 $\boldsymbol{Z}\in\mathbf{R}^{N\times D}$，$D$ 为随机特征维度。将 $\boldsymbol{Z}$ 按行随机划分为 B 块[18]，每个分块的样本规模为 $|B_j|$。随机傅里叶特征映射显式地构造随机特征空间，可在该随机特征空间中应用线性 SVM 来一致逼近高斯核 SVM[7]。

损失函数 $\max(1-y_i\boldsymbol{\omega}^{\mathrm{T}}z_i,0)^2$ 是 L-Lipschiz 连续的[20]。由分块经验风险最小化，可得

$$\boldsymbol{\omega}_B=\underset{\boldsymbol{\omega}_j}{\arg\min}\sum_{j=1}^{B}\sum_{i\in B_j}\max(1-y_i\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{z}_i,0)^2$$

此时，D-ADMM 中 $g(\cdot)$ 为指示函数[21]。定义为

$$\tilde{g}_S(x)=\begin{cases}0 & x\in S\\ \infty & x\notin S\end{cases}$$

则 $\boldsymbol{v}^{k+1}:=\Pi_S(\bar{\boldsymbol{\omega}}^{k+1}+\bar{\boldsymbol{u}}^k)$，其中 Π_S 表示在封闭凸集 S 上的投影。

大规模 SVM 最优分块数选择为

$$\hat{B}=\underset{B\in\mathcal{B}}{\arg\min}\sum_{j=1}^{B}\sum_{i\in B_j}\max(1-y_i\boldsymbol{\omega}_B^{\mathrm{T}}\boldsymbol{z}_i,0)^2+\eta E(B)$$

ADMM 框架下随机傅里叶特征空间中数据分块数选择过程见算法 2。

算法 2 分块数选择算法（SNC）

输入：随机特征矩阵 $\boldsymbol{Z}\in\mathbf{R}^{N\times D}$，分块数备选集 $\boldsymbol{B}$，且 $|\boldsymbol{B}|=t$，惩罚系数 η，并行效率下界 δ。

(1) 初始化：$E(B)=1$；

(2) for $k=1,2,\cdots,t$　do

(3) 调用 D-ADMM 计算 $\boldsymbol{\omega}_B=\arg\min_{\boldsymbol{\omega}_j}\sum_{j=1}^{B}\sum_{i\in B_j}\max(1-y_i\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{z}_i,0)^2$;

(4) 更新 $E(B)$;

(5) 计算 $V_k=\sum_{j=1}^{B}\sum_{i\in B_j}\max(1-y_i\boldsymbol{\omega}_B^{\mathrm{T}}\boldsymbol{z}_i,0)^2+\eta E(B)$;

(6) end for

(7) $\hat{B}=\arg\min_{B\in\boldsymbol{B}}V_k$。

输出:最优分块数 $\hat{B}$。

对 D-ADMM 中的并行子问题利用对偶坐标下降算法[10,22]求解。要得到 ε 精确解,迭代次数为 $O(\log(1/\varepsilon))$,时间复杂度为 $O(nD\log(1/\varepsilon))$,其中,n 为子数据块数据规模。算法 SNC 的迭代次数为 t,所以总的时间复杂度为 $O(tnD\log(1/\varepsilon))$。

损失函数 $\max(1-y_i\boldsymbol{\omega}^{\mathrm{T}}z_i,0)^2$ 和正则化函数 $1/2\|\boldsymbol{\omega}\|_2^2$ 均为封闭、适定凸函数,约束条件为仿射函数。根据 Slater 条件,对偶间隙为 0,强对偶性成立。故应用 D-ADMM 来并行/分布式求解大规模 SVM 满足收敛条件。

D-ADMM 框架下大规模 SVM 问题为

$$\min_{\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_{\hat{B}},\boldsymbol{o}} C\sum_{j=1}^{B}\sum_{i\in B_j}\max(1-y_i\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{z}_i,0)^2+\frac{1}{2}\|\boldsymbol{o}\|_2^2,\quad \text{s.t.}\quad \boldsymbol{\omega}_j-\boldsymbol{o}=\boldsymbol{0}$$

其中:C 为超参数,$\boldsymbol{o}$ 为全局变量,$\boldsymbol{\omega}_j$ 是与第 j 个进程的局部变量。由 D-ADMM 可得

$$\boldsymbol{\omega}_j^{k+1}=\arg\min_{\boldsymbol{\omega}_j}C\sum_{i\in B_j}\max(1-y_i\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{z}_i,0)^2+\frac{\rho}{2}\|\boldsymbol{\omega}_j-\boldsymbol{o}^k+\boldsymbol{u}_j^k\|_2^2$$

$$\boldsymbol{o}^{k+1}=\frac{\sum_{j=1}^{B}(\boldsymbol{\omega}_j^{k+1}+\boldsymbol{u}_j^k)}{\hat{B}+1/\rho}$$

$$\boldsymbol{u}_j^{k+1}=\boldsymbol{u}_j^k+\boldsymbol{\omega}_j^{k+1}-\boldsymbol{o}^{k+1}\qquad j=1,\cdots,\hat{B}$$

每个进程 j 处理一个子问题,各个并行子问题利用对偶坐标下降算法求解 $\boldsymbol{\omega}_j$[22]。

5　实验结果及分析

本节实现并行/分布式 ADMM 框架下随机傅里叶特征空间中的大规模 SVM,并实验验证所提出的数据分块数选择准则。实验中使用 6 个标准数据集,如表 1 所示,其中,大规模 SVM 的超参数 C 和高斯核参数 γ 通过在 11×11 空间内进行格搜索的 5-折交叉验证来选取,$(C,\gamma)\in\{2^{-9},2^{-7},\cdots,2^{7},2^{9}\}$。$\eta$ 为惩罚系数,并行效率下界 δ 设为 0.25,D-ADMM 中惩罚系数 ρ 设为 1,绝对误差 $\varepsilon^{\mathrm{abs}}$ 和相对误差 $\varepsilon^{\mathrm{rel}}$ 均设为 10^{-4},ε 设为 10^{-3}。实验环境:曙光"星云"高性能计算集群。采用 OpenMPI 1.4.5 和 C++实现并行算法。普通队列最多申请 4 节点,每个节点 32 个核,每个核分配内存 2 GB,主频 2.2 GHz,操作系统

表 1　标准数据集及相关参数

Tab. 1　Specification of benchmark datasets and related parameters

数据集	训练集	测试集	维度	η	$\log(C)$	$\log(\gamma)$	D
a9a	32 561	16 281	123	2.25	0	−6	1 000
ijcnn1	49 990	91 701	22	2.25	0	−8	500
w8a	49 749	14 951	300	2.25	2	−4	2 000
webspam	262 500	87 500	254	1.75	7	−8	1 000
covtype	435 759	145 253	54	2.25	6	−6	500
SUSY	3 750 000	1 250 000	18	3.00	2	−8	50

CentOS 5.8,作业管理系统 Torque 4.1.5。

对比实验结果如表 2 所示。其中，Acc 表示测试精度，计算时间和测试精度为重复 10 次实验的平均值。

表 2 最优分块数目、其他分块数目下并行计算时间(训练＋测试)与测试精度比较

Tab. 2 Comparison of parallel computation time (train ＋ test) and test accuracy (Acc) of optimal and other blocks

数据集	时间 /s	Acc/%	时间/s	Acc/%	时间/s	Acc/%	时间/s	Acc/%
a9a		$B=2$		$B=4$		$\hat{B}=6$		$B=12$
	19.0	85.17±0.03	14.2	85.21±0.02	10.8	85.21±0.03	12.8	85.17±0.02
ijcnn1		$B=2$		$\hat{B}=8$		$B=12$		$B=16$
	16.5	91.49±0.27	9.2	91.32±0.34	10.7	91.24±0.29	11.3	91.17±0.34
w8a		$B=2$		$\hat{B}=8$		$B=12$		$B=14$
	37.4	98.87±0.23	14.6	98.74±0.26	18.5	98.68±0.19	22.8	98.64±0.21
webpam		$B=2$		$B=6$		$\hat{B}=12$		$B=16$
	1 657	92.79±0.19	542	92.86±0.29	487	92.82±0.23	508	92.75±0.25
covtype		$B=2$		$B=6$		$\hat{B}=8$		$B=14$
	1 372	78.76±0.56	568	79.10±0.25	292	79.11±0.63	327	78.64±0.54
SUSY		$B=2$		$B=4$		$\hat{B}=10$		$B=14$
	270	79.16±0.03	162	79.16±0.03	65	79.15±0.02	101	79.15±0.03

实验结果表明,最优分块数相比于其他分块数下的并行模型预测精度相当,但计算时间大幅缩减。如在 webspam 数据集上，在最优分块数为 12,相比于 $B=2$ 而言,计算时间大幅缩短。

分块数过多,单节点任务规模太小,将会导致额外开销(如负载不均衡、进程间通信)增大,总时间反而会增加。如在 covtype 数据集上,在最优分块数为 $\hat{B}=8$ 下模型分类精度高于 $B=14$,同时计算时间更短;同样,在其他数据集上也可得出类似的结论。

为了验证不同分块数对模型测试精度的影响,给出测试精度随着迭代次数的变化曲线,如图 1 所示。

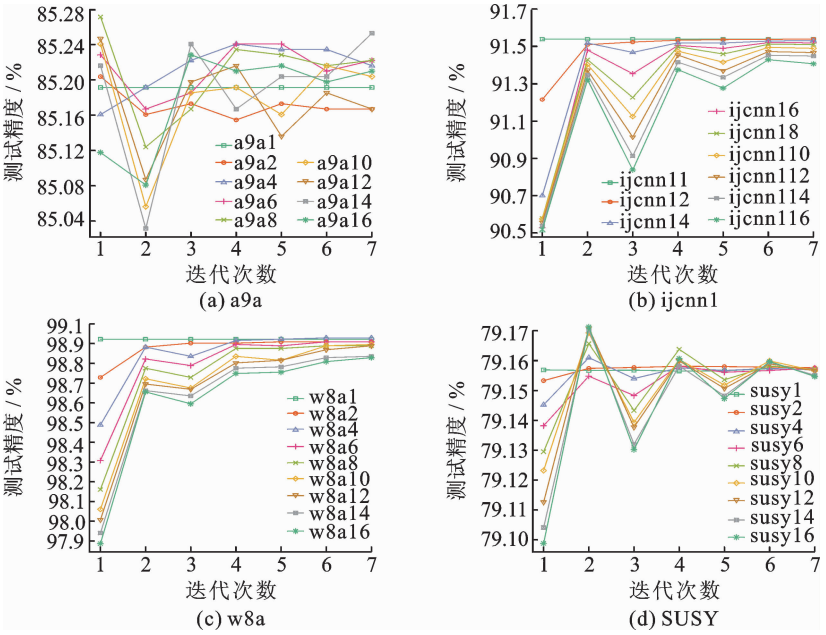

图 1 测试精度随着迭代次数的变化情况

Fig. 1 Test accuracy varies with respect to the number of iterations

结果表明,在相同迭代次数下,随着分块数的增多,模型的测试精度逐渐下降。这与第 2 节的泛化误差分析结果一致。

6 结束语

现有并行/分布式机器学习方法缺少有理论依据的数据分块数选择准则。针对这一问题,推导并行/分布式机器学习模型的泛化误差与分块数目的关系,折衷泛化性与并行效率,进而提出一个并行效率敏感的并行/分布式机器学习数据分块数选择准则。大规模支持向量机的理论分析和实验结果表明,所提出的数据分块数选择准则,可保证测试精度并提高计算效率。虽然所提出的数据分块数选择准则适用于 ADMM 框架下随机傅里叶特征空间中的大规模支持向量机,该数据分块数准则及分析方法也适用于其他并行/分布式机器学习模型,如大规模核岭回归等。

参考文献:

[1] Schölkopf B, Smola A J. Learning with kernels: Support vector machines, regularization, optimization, and beyond [M]. Cambridge, MA: MIT Press, 2002.

[2] 丁立中, 廖士中. 基于正则化路径的支持向量机近似模型选择[J]. 计算机研究与发展, 2012, 49(6):1248-1255.
Ding Lizhong, Liao Shizhong. Approximate model selection on regularization path for support vector machines[J]. Journal of Computer Research and Development, 2012, 49(6):1248-1255.

[3] Dong J X, Krzyzak A, Suen C Y. Fast SVM training algorithm with decomposition on very large data sets [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(4):603-618.

[4] Rahimi A, Recht B. Random features for large-scale kernel machines [C]//Advances in Neural Information Processing Systems. [S. l.]:[s. n.],2007:1177-1184.

[5] Rahimi A, Recht B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning [C]//Advances in Neural Information Processing Systems. [S. l.]:[s. n.],2009:1313-1320.

[6] Feng C, Hu Q, Liao S. Random feature mapping with signed circulant matrix projection [C]//Proceedings of the 24th International Joint Conference on Artificial Intelligence. [S. l.]:[s. n.],2015:3490-3496.

[7] 冯昌, 廖士中. 随机傅里叶特征空间中高斯核支持向量机模型选择[J]. 计算机研究与发展, 2016, 53(9):1971-1978.
Feng Chang, Liao Shizhong. Model selection for Gaussion kernel support vector machines in random Fourier feature space [J]. Journal of Computer Research and Development, 2016, 53(9):1971-1978.

[8] 王文剑, 祁晓博, 郭虎升. 一种高斯区间核 SVM 分类模型[J]. 数据采集与处理, 2017,32(1):46-53.
Wang Wenjian, Qi Xiaobo, Guo Husheng. Support vector machine classification model based on Gaussian interval kernel[J]. Journal of Data Acquisition and Processing, 2017,32(1):46-53.

[9] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers [J]. Foundations and Trends in Machine Learning, 2011, 3(1):1-122.

[10] Zhang C, Lee H, Shin K G. Efficient distributed linear classification algorithms via the alternating direction method of multipliers[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. [S. l.]:[s. n.],2012:1398-1406.

[11] Avron H, Sindhwani V. High-performance kernel machines with implicit distributed optimization and randomization [J]. Technometrics, 2016, 58(3):341-349.

[12] Yu Z Q, Shi X J, Yan L, et al. Distributed stochastic ADMM for matrix factorization [C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. [S. l.]:[s. n.],2014:1259-1268.

[13] Zhang Y, Duchi J, Wainwright M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates[J]. Journal of Machine Learning Research, 2015, 16:3299-3340.

[14] Deng W, Lai M J, Peng Z, et al. Parallel multi-block ADMM with $o(1/k)$ convergence [J]. Journal of Scientific Computing, 2016:1-25.

[15] Nishihara R, Lessard L, Recht B, et al. A general analysis of the convergence of ADMM [C]//Proceedings of the 32nd In-

ternational Conference on Machine Learning. [S. l.]: [s. n.], 2015: 343-352.

[16] Xu C, Zhang Y, Li R, et al. On the feasibility of distributed kernel regression for big data [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(11): 3041-3052.

[17] Vapnik V N, Vapnik V. Statistical learning theory [M]. New York: Wiley & Sons, 1998.

[18] Yu H F, Hsieh C J, Chang K W, et al. Large linear classification when data cannot fit in memory [C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l.]: [s. n.], 2010: 833-842.

[19] Bertsekas D P, Tsitsiklis J N. Parallel and distributed computation: Numerical methods [M]. Englewood Cliffs, New Jersey: Prentice-Hall, 1989.

[20] Rosasco L, De Vito E, Caponnetto A, et al. Are loss functions all the same? [J]. Neural Computation, 2004, 16(5): 1063-1076.

[21] Boyd S, Vandenberghe L. Convex optimization [M]. Cambridge: Cambridge University Press, 2004.

[22] Hsieh C J, Chang K W, Lin C J, et al. A dual coordinate descent method for large-scale linear SVM [C]//Proceedings of the 25th International Conference on Machine Learning. [S. l.]: [s. n.], 2008: 408-415.

作者简介:

张闯(1991-),男,硕士研究生,研究方向:并行/分布式机器学习、模型选择,E-mail: c_zhang@tju.edu.cn。

廖士中(1964-),男,博士,教授,博士生导师,CCF 会员,研究方向:人工智能应用基础、理论计算机科学。

(编辑:夏道家)