一种改进的贝叶斯逻辑回归核心集构建算法

张士翔 李汪根 李 童 朱楠楠
(安徽师范大学计算机与信息学院 安徽 芜湖 241000)

摘 要 随着互联网的高速发展,新型信息发布方式不断涌现,由此所产生的数据正以前所未有的速度"爆炸式"增长。如何处理和分析庞大的原始数据,并将之变成可用知识加以学习和利用,已成为国内外科学家和技术专家共同关注的重要课题。贝叶斯方法提供了丰富的分层模型、不确定的量化及预先的规范,因此其在大规模数据背景下的使用十分具有吸引力。限制迭代的二分 K-means 算法保留了近似标准二分 K-means 算法的聚类质量且拥有更高的计算效率,更适用于需要处理速度更快的大型数据集。针对原有核心集构建算法执行效率低的问题,对限制迭代的二分 k-means 算法进行改进,使其在保证聚类效果的情况下更快速地得到聚类结果并计算相关数据点权值,从而构建出核心集。实验证明,与原算法相比,改进后算法的计算效率更高,近似性能相近且在部分情况下近似效果更优。

关键词 核心集,限制迭代二分 k-means,贝叶斯逻辑回归,大规模数据集
中图法分类号 TP391 **文献标识码** A

Improved CoreSets Construction Algorithm for Bayesian Logistic Regression

ZHANG Shi-xiang LI Wang-geng LI Tong ZHU Nan-nan
(School of Computer and Information, Anhui Normal University, Wuhu, Anhui 241000, China)

Abstract With the rapid development of the Internet, new types of information dissemination methods are emerging. It leads to an explosion of data at an unprecedented rate. How to process and analyze huge raw data and turn it into usable knowledge for learning and utilization, has become an important topic of common concern for scientists and technical experts at home and abroad. The Bayesian approach provides rich hierarchical models, uncertainty quantification and prior specification, so in large-scale data background it is very attractive. The limited-iteration bisecting K-means algorithm preserves the clustering quality of the approximate standard bisecting K-means algorithm with higher computational efficiency, and it is more suitable for large data sets requiring faster processing speeds. Aiming at the low execution efficiency problem of the original coresets construction algorithm, the limited-iteration bisecting K-means algorithm is improved, making the clustering result obtained at a faster speed and the weight of the relevant data points calculated under the condition of ensuring the clustering effect, thus constructing the coresets. Experiments show that compared with the original algorithm, the improved algorithm has higher computational efficiency, similar approximation performance and better approximation effect in some cases.

Keywords Coresets, Limited-iteration bisecting K-means, Bayesian logistic regression, Large-scale dataset

1 引言

随着互联网的高速发展,数据正以前所未有的速度积累,越来越多的科学家与商业人士开始使用大规模的数据集进行科学研究与商业分析。在这样的规模下,即使是一些简单的操作,例如检查每个数据点,整体任务也会变得十分繁重。同时,这些任务所需数据也难以存储在单台机器的物理内存中,这在过去限制了从业者采用相对简单的统计建模方法。然而,贝叶斯方法具有丰富的分层模型,不确定性量化和预先的规范,使得贝叶斯推断过程可以扩展到大规模数据环境。

大规模数据的贝叶斯推理的常用方法是修改特定的推演算法,从而适应流式或分布式处理环境。例如基于变分贝叶斯的子采样方法[1-2]、基于马尔可夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)的子采样方法[3-4]以及基于 MCMC 的分布式"一致性"方法[5-6],然而,这些方法都有一定的实践和理论限制。随机变分推理[7]和子采样 MCMC 方法每次迭代使用一个新的随机数据子集,这需要对数据进行随机访问,因此其不能用于无法完全装入内存的庞大的数据集。此外,在实践中发现子采样 MCMC 方法每次迭代需要检查数据的一个恒定部分,这严重限制了计算增益[8-9]。诸如一致性 MCMC[10-11]和流变分贝叶斯等方法计算效率的提高缺乏验证且不能保证推演结果的质量。Huggins 等[12]利用数据冗余的概念提出了一种可修改的贝叶斯推理框架,从数据集角度出发,构建一个近似替代原数据集的加权子集,称其为核心集。文中对构建核心集的过程给出了严格的理论证明,但是该算法本身须对数据集进行 K-means 聚类,在大规模数据集

本文受国家自然科学基金(111023),高校领军人才引进与培育计划项目(051619)资助。
张士翔(1991—),男,硕士,主要研究方向为机器学习;**李汪根**(1973—),男,博士,教授,主要研究方向为 DNA 计算、智能计算和模式识别等,E-mail:xchen@mail.ahnu.edu.cn(通信作者)。

下计算效率有待提高。

二分 K-means 是一种标准 K-means 的改进算法，在聚类质量和效率方面较标准 K-means 均有提升。Steinbach 等[13]提出了二分 K-means 算法，其使用 Lloyd 二均值(即 K 值为 2 的 K-means)首先将数据集划分为两个群集，产生两个聚类。然后在群集总数达到 K 之前，选择当前群集池中的群集，将其划分为两个群集，并将群集总数加 1。该等分过程继续进行，直到聚类总数达到 K。为了使群集平分，在选择相同的群集上执行多个二均值运算，并选择最佳平分结果。为了进一步增强聚类结果，可以使用标准 K-means 来细化从二等分过程获得的 K 个聚类，其中将二等分过程产生的 K 个聚类的质心作为标准 K-means 的 K 个初始聚类中心。文献[13]的研究表明，一般的二分 K-means 较具有随机初始聚类中心的标准 K-means 拥有更好和更一致的聚类结果，并且细化时聚类结果提升更明显。自提出以来，众多研究人员一直致力于在各种应用中使用二分 K-means 以及在算法上对二分 K-means 进行改进。Savaresi 等[14]研究了两对初始中心的标准 K-means($K=2$)和使用此初始中心的标准 K-means($K=2$)的收敛特性。Liu 等[15]提出了一种基于一对初始中心的标准 K-means($K=2$)的二分 K-means 算法，并且该算法每次只平分一个聚类簇；还研究了一种新技术，用于精炼等分过程产生的团簇。与没有细化的二分 K-means 相比，Liu 等的精细二分 K-means 具有更好的聚类质量。Zhuang 等[16]提出了一种限制迭代的二分 K-means(Limit-Iteration Bisecting K-means，LIBKM)算法，该算法在维持了标准 K-means 算法的聚类质量的基础上具有更高的计算效率，适用于大规模数据集。

基于上述分析，本文提出了一种改进的贝叶斯逻辑回归核心集构建算法(Improved CoreSets Construction Algorithm for Bayesian LogisticRegression，ICCAFBLR)。在构建核心集过程中，通过使用改进的限制迭代二分 K-means，提高了整体算法的执行效率。同时，根据原有核心集构建算法的严格理论证明，保证了整体算法的可靠性及核心集的近似效果。

2　相关工作

2.1　相关记号

根据贝叶斯推理的一般定义，$\boldsymbol{D}=\{(X_n,Y_n)\}_{n=1}^{N}$ 代表整个数据集，其中 $X_n\in\boldsymbol{X}$ 是协变量的向量，$Y_n\in y$ 是一个目标值。$\pi_0(\theta)$ 是基于 $\theta(\theta\in\Theta)$ 的先验密度，$p(Y_n\mid X_n,\theta)$ 是给定参数 θ 估计 n 个目标值的似然。

2.2　可拓展的贝叶斯对数几率回归核心集算法

大规模数据环境中的重要见解是：除去少部分独特的稀有数据外，大部分数据往往是多余的。可拓展的贝叶斯对数几率回归核心集算法是利用数据冗余开发的可修改的贝叶斯推理框架，可以被认为是一个预处理步骤。该方法通过构建核心集(数据的加权子集)以达到近似于整个数据集的推理效果，而不是通过修改推理算法的常见做法。

贝叶斯后验密度如下所示：

$$\pi_N(\theta):=\frac{\exp(\boldsymbol{L}_N(\theta)\pi_0(\theta))}{\varepsilon_N} \tag{1}$$

其中，$\boldsymbol{L}_N(\theta):=\sum_{n=1}^{N}\ln p(Y_n\mid X_n,\theta)$ 是模型的对数似然，$\varepsilon_N:=\int\exp(L_N(\theta))\pi_0(\theta)\mathrm{d}\theta$ 是边际似然。旨在构建一个加权的数据子集 $\widetilde{\boldsymbol{D}}=\{(\gamma_m,\widetilde{X}_m,\widetilde{Y}_m)\}_{m=1}^{M}$，其中 $M\ll N$，加权对数似然定义为 $\widetilde{\boldsymbol{L}}_N(\theta):=\sum_{m=1}^{M}\gamma_m\ln p(\widetilde{Y}_m\mid\widetilde{X}_m,\theta)$，满足以下条件：

$$|\boldsymbol{L}_N(\theta)-\widetilde{\boldsymbol{L}}_N(\theta)|\leqslant\varepsilon\boldsymbol{L}_N(\theta),\forall\theta\in\Theta \tag{2}$$

满足式(2)的 $\widetilde{\boldsymbol{D}}$ 是总数据集 $\boldsymbol{D}$ 的一个 ε 核心集，近似的后验概率为：

$$\widetilde{\pi}_N(\theta):=\frac{\exp(\widetilde{\boldsymbol{L}}_N(\theta)\pi_0(\theta))}{\widetilde{\boldsymbol{\varepsilon}}_N} \tag{3}$$

其中，近似于真实边际似然 $\boldsymbol{\varepsilon}_N$ 的 $\widetilde{\boldsymbol{\varepsilon}}_N$ 定义为：

$$\widetilde{\boldsymbol{\varepsilon}}_N:=\int\exp(\widetilde{\boldsymbol{L}}_N(\theta))\pi_0(\theta)\mathrm{d}\theta \tag{4}$$

正如命题 1 所示，在贝叶斯估计方面，ε 核心集是一个有用的近似概念。

命题 1　设 $\boldsymbol{L}_N(\theta)$ 和 $\widetilde{\boldsymbol{L}}_N(\theta)$ 为任意的非正对数似然函数并且满足 $|\boldsymbol{L}_N(\theta)-\widetilde{\boldsymbol{L}}_N(\theta)|\leqslant\varepsilon\boldsymbol{L}_N(\theta)$，在所有的 $\theta\in\Theta$ 的情况下，对于任何的先验密度 $\pi_0(\theta)$，$\boldsymbol{\varepsilon}_N:=\int\exp(L_N(\theta))\pi_0(\theta)\mathrm{d}\theta$ 和 $\widetilde{\boldsymbol{\varepsilon}}_N:=\int\exp(\widetilde{\boldsymbol{L}}_N(\theta))\pi_0(\theta)\mathrm{d}\theta$ 是有限的，边际似然满足条件：

$$|\ln\boldsymbol{\varepsilon}-\ln\widetilde{\boldsymbol{\varepsilon}}|\leqslant\varepsilon|\ln\varepsilon| \tag{5}$$

在对数几率回归中，协变量是实特征向量 $X_n\in R^D$，目标值是标签 $Y_n\in\{-1,1\}$，$\Theta\in R^D$，似然函数定义为：

$$p(Y_n\mid X_n,\theta)=p_{logistic}(Y_n\mid X_n,\theta):=\frac{1}{1+\exp(-X_nY_n\cdot\theta)}$$

在整个工作分析中，高斯分布来定义先验密度 $\pi_0(\theta)$。为了表达简洁，定义：

$$Z_n:=X_nY_n\ \text{和}\ \emptyset(s):=\ln(1+\exp(-s)) \tag{6}$$

在计算上直接选择最优的 ε 核心集是不可行的，所以只有采取间接的方法。其利用一个被称为敏感度的量度 $\sigma_n(\Theta)$ 来进行核心集构建，并证明正确性。对于整个数据集上的任意点，越重要的点敏感度越高，冗余性越低。在对数几率回归的背景下，将敏感度定义为：

$$\sigma_n(\Theta):=\sup_{\theta\in\Theta}\frac{N\emptyset(Z_n\cdot\theta)}{\sum_{l=1}^{N}\emptyset(Z_l\cdot\theta)} \tag{7}$$

直观的说，$m_n\geqslant\sigma_n(\Theta)$ 代表了在变化参数 $\theta\in\Theta$ 的条件下每个样本点对于对数似然 $\boldsymbol{L}_N(\theta)$ 的影响程度，因此高敏感度的样本点应该被包含于核心集中。然而，由于不能直接计算 $\sigma_n(\Theta)$，不得不找到一个上界 $m_n\geqslant\sigma_n(\Theta)$。因此，最关键的问题变成了高效计算出每个样本点敏感度的紧上界。

考虑在任何 $R>0$ 的情况下令 $\Theta=B_R(\{B_R:=\theta\in R^D\mid\|\theta\|_2\leqslant R\})$。因为数据集被预处理成均值为 0 和方差为 1，所以选择一个基于欧几里德球型参数空间是合理的。

聚集的样本点通常是冗余的，而与其他样本点相距甚远的点对推演结果的影响很大，基于此，构建了敏感度上界。聚类是一种有效地总结数据和探测离群点的方式，所以其使用 k 聚类来构建敏感度上界。一个 k 聚类有 k 个聚类中心 $\boldsymbol{Q}=\{Q_1,\cdots,Q_K\}$，属于每个聚类的向量集合定义为 $G_i:=\{Z_n\mid i=\underset{j}{\operatorname{argmin}}\|Q_j-Z_n\|_2\}$，令 $G_i^{(-n)}:=G_i\backslash\{Z_n\}$。定义 $Z_{G,i}^{(-n)}$ 是来自于 $G_i^{(-n)}$ 的随机向量，而 $\overline{Z}_{G,i}^{(-n)}:=E|Z_{G,i}^{(-n)}|$ 是其均值。其使用 k 聚类来构建一个可快速计算的近似上界 $\sigma_n(B_R)$，其定义如下：

$$\sigma_n(B_R)\leqslant m_n:=\frac{N}{(1+\sum_{i=1}^{k}|G_i^{(-n)}|\mathrm{e}^{-R\|Z_{G,i}^{(-n)}-Z_{G,i}^{(-n)}\|_2}} \tag{8}$$

3 改进的贝叶斯逻辑回归核心集构建算法

原有的核心集算法采用 K-means 聚类，在大规模数据集下进行该聚类，效率不理想且影响整体效率。为了解决上述问题，本文对原有 LIBKM 算法进行改进后与核心集构建算法相结合，提出了一种改进的贝叶斯逻辑回归核心集构建算法，实验证明了与原核心集构建算法相比，该算法的近似效果基本不变，但提升了运行效率。

3.1 聚类中心的选择

对于聚类算法，错误的聚类中心会直接导致错误的样本分类。由于算法每次迭代只有两个初始聚类中心并且迭代次数是有限制的，为了保证聚类质量以及避免最糟糕的聚类结果，必须谨慎地选择该对聚类中心，使用限制迭代的平分集合二均值（The One-Iteration Set-Bisecting Two-Means，LISBTM）算法，具体算法步骤如下。

步骤 1 给定数据集 S；

步骤 2 计算需划分数据集的均值点 S；

步骤 3 计算所有样本点到均值 S 的距离，找到其中最远的点 P，即聚类中心点 $C1$；

步骤 4 令 $C2=2P-C1$，即另一聚类中心；

步骤 5 计算各样本点至 $C1$ 和 $C2$ 的距离，并划分至最近的聚类簇中，得到数据子集 $S1$ 和 $S2$；

步骤 6 重新计算 $C1$ 和 $C2$，分别为数据子集 $S1$ 和 $S2$ 的均值；

步骤 7 返回 $C1$ 和 $C2$ 及子集 $S1$ 和 $S2$。

3.2 改进的限制迭代二分 K-means 算法

原始的二分 K-means 算法在平分数据集时需要多次执行 K-means($K=2$)算法，其中的每次运行都以一对不同的初始中心开始。为了提高计算效率，改进的二分 K-means 算法在每次平分数据集时只执行一次 K-means 算法。为了保证聚类质量，该算法最后通过标准 K-means 对聚类结果进行微调。同时，为保证算法时间效率，迭代次数仅设为 3，具体算法步骤如下。

步骤 1 给定聚类中心个数 K，初始样本集为总样本集 S，初始聚类中心 C；

步骤 2 计算每个样本集聚类的平方误差和，选取其中最大的集合使用 LISBTM 算法，得到子集合 $S1$，$S2$ 及新子集合的聚类中心 $C1$，$C2$；

步骤 3 将 S 移出样本集并将 S1 和 S2 加入样本集，将 C 移出聚类中心集合，将 C1 和 C2 加入聚类中心集合；

步骤 4 若聚类中心个数小于 K，返回步骤 2，否则执行步骤 5；

步骤 5 在得到的聚类中心基础上执行标准 K-means 算法。

3.3 整体的算法流程

本文提出的 ICCAFBLR 首先根据 LISBTM 算法计算初始聚类中心；再根据 ILIBKM 算法执行，在原有 LIBKM 算法基础上对聚类结果进行小幅度修正；最后，根据聚类结果执行核心集构建算法，计算敏感度，得到每个样本点的权重，以得到所需的加权子集。ICCAFBLR 流程图如图 1 所示。

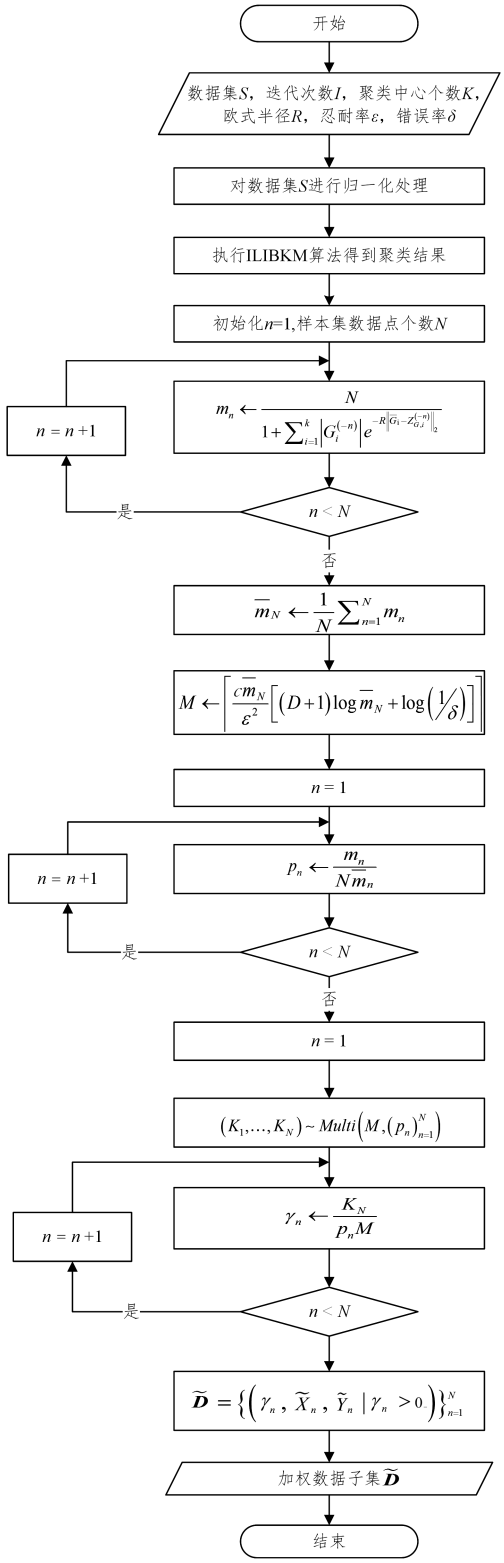


图 1 ICCAFBLR 流程图

算法的步骤如算法 1 所示。

算法 1 改进的贝叶斯对数几率回归核心集构建算法

输入：数据集 **D**，聚类中心个数 K，半径 R，容忍度>0，错误率

输出：核心集 ε

步骤 1 根据式(8)使用 ILIBKM 算法计算各数据集样本点的敏感度上界 m_n；

步骤 2 计算数据集 m_n 的均值 $\bar{m}_n=\frac{1}{N}\sum_{n=1}^{N} m_n$；

步骤 3 计算核心集的样本点个数 $M=\left\lceil\frac{C\bar{m}_n}{\varepsilon^2}\right.[(D+1)\log\bar{m}_n+$

$\log(1/\delta)\rceil$，其中 c 是一个常量；

步骤 4　计算各样本点敏感度权重 $p_n=\frac{m_n}{N\bar{m}_n}$；

步骤 5　由 $(K_1,\cdots,K_N)\sim \mathrm{Multi}(M,(p_n)_{n=1}^N)$ 计算各样本点核心集权重 $\gamma_n=\frac{K_N}{M p_n}$；

步骤 6　得到 ε-核心集 $\tilde{\mathbf{D}}=\{\gamma_n,X_n,Y_n\mid\gamma_n>0\}_{n=1}^N$。

4　实验及结果分析

本文实验主要针对大规模数据集下贝叶斯对数几率回归问题，为了验证算法的健壮性，分别使用可扩展的贝叶斯对数几率回归算法（Coresets for Scalable Bayesian Logistic Regression，CFSBLR）、随机采样算法（Random Subsampling，RS），本文提出的算法在真实和模拟数据集上进行对比。

实验运行环境为 Hasee PC，Inter(R) Core(TM) CPU i5-7400 @ 3.00 GHZ，8 GB 内存。算法使用 Python 语言编写。

4.1　实验数据集

模拟数据集（Binary10）：根据模型 $X_{nd}\overset{indep}{\sim}Bern(p_d)$，$d=1,\cdots,D(D=10)$）和 $Y_n\overset{indep}{\sim}p_{logistic}(\cdot\mid X_n,\theta)$ 来模拟生成二进制数据。模拟数据中存在少量很少发生但具有高度预测性的特征，这是一种常见的现实世界现象。因此，$p=(1,0.2,0.3,0.5,0.01,0.1,0.007,0.005,0.001)$，$\theta=(-3,1.2,-0.5,0.8,3,-1,-0.7,4,3.5,4.5)$。

真实数据集（ChemReact）：由 $N=26\,733$ 种化学物质组成，每种化学物质具有 $D=100$ 的特性，目标是预测每种化学物质是否具有反应性。

封面类型（CovType）数据集：由 $N=581\,012$ 个制图观察组成，具有 $D=54$ 个特征。任务是预测每个观察位置的树木类型。

本实验所用数据集的特征如表 1 所列。

表 1　实验数据集

NAME	N	D	Positive examples/%	k
Binary(10)	1M	10	8.9	4
ChemReact	26733	100	3	6
Covtype	581012	54	51	6

4.2　评价指标

实验主要使用了两个评价指标来评价核心集的近似性能，分别是：MMD（Maximum Mean Discrepancy）和 TLL（Test Log Likelihood），MMD 值越低 TLL 值越高表示核心集算法的近似效果越好。计算公式如下：

$$MMD=\left\|\frac{1}{n}\sum_{i=1}^{n}\emptyset(x_i)-\frac{1}{m}\sum_{j=1}^{m}\emptyset(y_j)\right\|_H^2 \tag{9}$$

MMD 主要用于度量两个不同但相关的分布的距离，其中 H 表示这个距离是由 $\emptyset()$ 将数据映射到再生希伯尔空间（RKHS）中进行度量的。

除此之外，TLL 表示样本集的测试对数似然，计算公式如下：

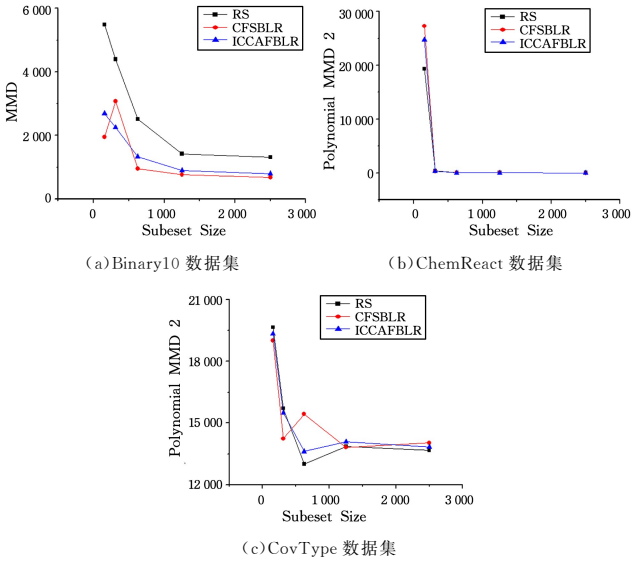
$$TLL=\frac{1}{N\times M}\sum_{i=1}^{N}\sum_{k=1}^{M}\log(1+\exp(Z_i\cdot\theta_k)) \tag{10}$$

其中，θ_k 是训练样本集的特征值向量，$Z_i:=X_i\ Y_i$，X_i 是测试集的样本特征集合，Y_i 是测试集的标签向量。

同时，对比各个算法构建替代子集所花费的时间，以比较各算法的运行效率。

4.3　结果分析

使用自适应 Metropolis 算法[17-18]进行后验推理。对于每个数据集，我们为每个子样本（$M=156,312,625,1\,250,2\,500$）迭代运行两种核心集和随机子采样算法 5 次。在完整数据集和每个子采样数据集上运行自适应 Metropolis 算法，迭代 500 次。启发式方法被用于选择尽可能大的 R，这仍然获得了适度的总灵敏度界限。较低维度的生成数据集使用 $k=4$，而更高维度的真实世界数据集使用 $k=6$。结果如图 2 所示。



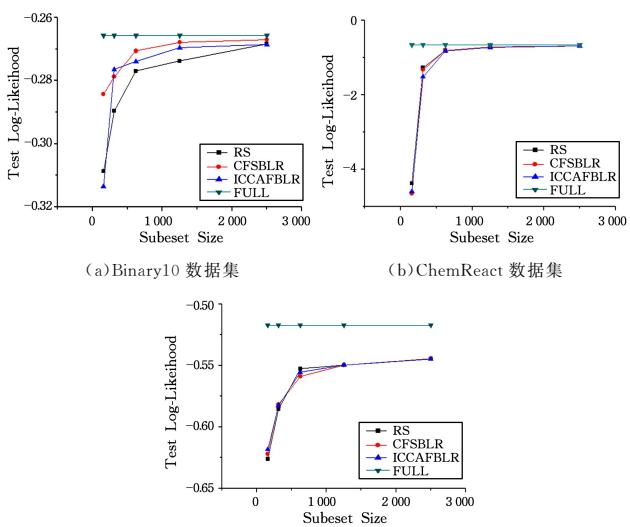
(a)Binary10 数据集　(b)ChemReact 数据集

(c)CovType 数据集

图 2　各算法 MMD 值的对比

通过图 2 可以发现，在各种情况下，本文所提的算法都体现了与原算法相当的近似性能，且对比图 2(a)可以看出，ICCAFBLR 和 CFSBLR 算法近似结果相近且明显优于 RS 算法。从图 2(c)中可以比较看出，ICCAFSBLR 算法性能较其余两种方法更加稳定，具有更好的鲁棒性。

为了更好地验证本文所提算法的性能，图 3 给出了各算法间 TLL 值的变化对比。在模拟数据集 Binary10 上，本算法性能近似于原算法，而在真实数据集上，改进算法的性能均要优于原核心集。

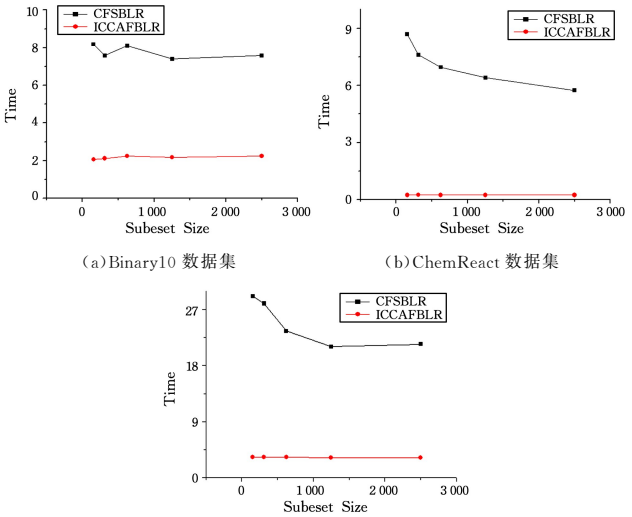


(a)Binary10 数据集　(b)ChemReact 数据集

(c)CovType 数据集

图 3　各算法 TTL 值的变化对比

核心集构建时间占总推演算法时间比例如图 4 所示。



(a)Binary10 数据集 (b)ChemReact 数据集

(c)CovType 数据集

图 4 各算法构建核心集的时间对比

如图 4 所示,在所有测试数据集下 ICCAFBLR 算法构建核心集的时间基本稳定,运行时间与构建核心集的大小无关。同时,在所有数据集上,ICCAFSBLR 算法构建核心集的速度均优于 CFSBLR 算法。

结束语 针对大规模数据集下核心集构建效率低的问题,本文提出了一种改进的可拓展贝叶斯逻辑回归的核心集构建算法,将限制迭代的二分 K-means 算法与核心集构建算法相结合并加以改进。在许多情况下,该算法能够更快地获得高质量的逻辑回归后验近似。但是,本文算法中的参数如聚类中心个数 K 等是事先预设的特定值,不能自动生成,因此如何根据不同的数据集自适应的生成所需的参数聚类中心数等是一个问题。同时,该核心集算法仅针对贝叶斯回归问题,下一步研究的重点是对其加以改进以适用于其他算法。

参 考 文 献

[1] BRODERICK T,BOYD N,WIBISONO A,et al. Streaming Variational Bayes[C]// Advances In Neural Information Processing Systems. Necada,USA:MIT Press,2013:1727-1735.

[2] CAMPBELL T,STRAUB J,III J W F,et al. Streaming,Distributed Variational Inference for Bayesian Nonparametrics[C]// Advances in Neural Information Processing Systems. Montreal,Canada:MIT Press,2015:280-288.

[3] AHN S,KORATTIKARA A,WELLING M. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. [C]// International Conference on Machine Learning. Edinburgh, Scotland: ACM,2012:1591-1598.

[4] BARDENET R,DOUCET A,HOLMES C. Towards scaling up Markov chain Monte Carlo:An adaptive subsampling approach [C]// International Conference on Machine Learning. Beijing: ACM,2014:405-413.

[5] ENTEZARI R,CRAIU R V,ROSENTHAL J S. Likelihood inflating sampling algorithm[J]. Canadian Journal of Statistics, 2017,46:147-175.

[6] RABINOVICH M,ANGELINO E,JORDAN M I. Variational consensus Monte Carlo[C]// Advances in Neural Information Processing Systems. Montreal,Canada:MIT Press,2015:1207-1215

[7] HOFFMAN M D,BLEI D M,WANG C,et al. Stochastic variational inference[J]. Journal of Machine Learning Research, 2013,14(1):1303-1347.

[8] ALQUIER P,FRIEL N,EVERITT R,et al. Noisy Monte Carlo:convergence of Markov chains with approximate transition kernels[J]. Statistics and Computing,2016,26(1/2):29-47.

[9] BARDENET R,DOUCET A,HOLMES C. On Markov chain Monte Carlo methods for tall data[J]. Journal of Machine Learning Research,2016,18:1-43.

[10] SCOTT S L,BLOCKER A W,BONASSI F V,et al. Bayes and big data:the consensus Monte Carlo algorithm[J]. International Journal of Management Science and Engineering Management, 2016,11(2):78-88.

[11] SRIVASTAVA S,CEVHER V,DINH Q,et al. WASP:Scalable Bayes via barycenters of subset posteriors[C]// Proceedings of the International Conference on Artificial Intelligence and Statistics. San Diego,California,USA:JMLR,2015:912-920.

[12] HUGGINS J H,CAMPBELL T,BRODERICK T. Coresets for Scalable Bayesian Logistic Regression[C]// Advances in Neural Information Processing Systems. Barcelona,Spain:MIT Press, 2016:4080-4088.

[13] STEINBACH M,KARYPIS G,KUMAR V,et al. A comparison of document clustering techniques[C]// KDD Workshop on Text Mining. Boston,USA:2000:525-526

[14] SAVARESI S M,BOLEY D L. On the Performance of Bisecting K-Means and PDDP[J]. Intelligent Data Analysis,2004,8(4): 345-362.

[15] LIU G C,HUANG T T,CHEN H N. Improved Bisecting K-means Clustering Algorithm[J]. Computer Application and Software,2015,32(2):261-263.

[16] ZHUANG Y,MAO Y,CHEN X. A Limited-Iteration Bisecting K-Means for Fast Clustering Large Datasets[C]// IEEE Trust Com-Big Data SE-ISPA. Tianjin,China:IEEE,2017:2257-2262.

[17] HAARIO H. An Adaptive Metropolis Algorithm[J]. Bernoulli, 2001,7(2):223-242.

[18] ROBERTS G O,TWEEDIE R L. Exponential Convergence of Langevin Distributions and Their Discrete Approximations[J]. Bernoulli,1996,2(4):341-363.