高维数据挖掘中基于中位数回归的特征提取新方法*

李泽安[a], 陈建平[a], 赵为华[b]
(南通大学 a. 计算机科学与技术学院; b. 理学院, 江苏 南通 226019)

摘 要: 为降低噪声对数据特征提取(变量选择)效果的不利影响,基于中位数回归分析方法,利用变量选择降维技术(正则化估计),提出了一种稳健、有效的特征提取(变量选择)新方法,并具体给出了估计算法,该算法具有快速计算的特点。实验结果表明,新方法能够有效地对高维数据集进行估计和变量选择,且具有较高的准确性,即使数据中的信噪比很低时,该方法仍具有较好的效果。因此,该方法为高维数据挖掘特征提取提供了稳健且有效的方法。

关键词: 高维数据; 特征提取; 变量选择; 中位数回归; LASSO
中图分类号: TP312 **文献标志码**: A **文章编号**: 1001-3695(2013)02-0374-03
doi:10.3969/j.issn.1001-3695.2013.02.014

New feature extraction method in high-dimensional data mining based on median regression

LI Ze-an[a], CHEN Jian-ping[a], ZHAO Wei-hua[b]
(*a. School of Computer Science, b. School of Science, Nantong University, Nantong Jiangsu* 226019, *China*)

Abstract: In order to reduce the unfavorable influence of nosie existing to the feature extraction(variable selection), this paper proposed a robust and effictive method based on median regression and dimensional reduction technique of variable selection(regularized estimation). Moreover, it gave the detail of algorithm which possessed the advantage of fast computation. Simulations show that the new method can effictively estimate and select important variables with high correctness. Even in the case of very low signal to noise ratio, this method still has good result compared with other method. It is really an efficient and robust feature extraction method for high-dimensional data mining.

Key words: high-dimensional data; feature extraction; variable selection; median regression; LASSO

高维数据挖掘[1,2]和传统的数据挖掘主要差别在于高维度,即数据中变量的个数远远大于数据的观测数目。高维数据在各领域经常出现,如基因芯片数据、基因表达数据、多媒体数据、Web 使用数据、时间序列数据和医学影像数据等,它们的维度通常可以高达到成百上千。由于维数祸根的影响,使得高维数据的挖掘变得异常困难,同时也使得高维数据挖掘成为数据挖掘领域中的前沿和热点。高维数据的一个共同特征往往是真实的影响变量并不太多,许多变量是冗余的,即高维数据具有变量的稀疏性。比如高光谱的卫星图像维数为上万维,但是用户感兴趣的区域在整个图像当中所占比例非常之小,因此如果利用稀疏性特征,就可以从高维海量数据当中提取出真实有意义的特征来。如何从成百上千维变量中有效地选择出真实的影响变量,从而达到变量选择,提高推断和预测精度的目的是进行高维数据挖掘需要解决的问题。最基本的想法是通过降维将数据从高维降到低维,然后用低维数据的处理办法进行处理。

回归分析是数据挖掘技术中重要的方法[1],它能揭示因变量与自变量之间的关系,从而研究数据挖掘问题中客观存在的规律,并作出合理的预测分析。通过对数据进行预分析,建立恰当的回归模型,能将数据从高维降到低维。线性回归模型是最常用的数据挖掘模型,其目的就是用一个或多个自变量的变化去解释因变量的变化,通过检验模型、估计预测等环节找出自变量与因变量的关系,挖掘出实际问题中的有用信息,为进一步决策提供科学依据。

另一方面,为了减小可能存在的模型误差,初始回归建模时,往往会引入很多可能与之相关的变量。为了提高模型的预测精度,增强模型的可解释性,就需要选择对因变量有显著影响的重要解释变量。因此,变量选择(或称特征提取)是数据进行有效数据挖掘的一个很重要的步骤,也成为当今数据挖掘研究中的热点问题[2]。然而,实际问题中收集到的数据往往具有异质性,数据中具有较大的噪声,即信噪比不高,为实际工作者进行数据特征提取带来了很大的麻烦。本文将研究数据挖掘中稳健且有效的特征提取方法,该方法能不受或少受数据噪声的影响,正确地选择出重要变量,进而实现很好的数据挖掘效果。

1 特征提取(变量选择)的研究现状

1.1 最小二乘估计

经典的线性回归模型可表示为

收稿日期: 2012-06-30; **修回日期**: 2012-08-24 **基金项目**: 江苏省自然科学基金资助项目(SBK200920379);南通大学自然科学基金资助项目(10Z008)

作者简介:李泽安(1977-),女,讲师,主要研究方向为数据挖掘、统计学习(li.za@ntu.edu.cn);陈建平(1963-),男,院长,教授,主要研究方向为人工智能;赵为华(1979-),男,讲师,博士,主要研究方向为统计分析.

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad i = 1, \cdots, n \tag{1}$$

其中:y_i 是因变量;x_{ij} 是解释变量;μ 是常数项;$\beta_j (j = 1, \cdots, p)$ 是回归系数;ε_i 是模型误差,其均值为 0,方差为 σ^2。

对于式(1),首要问题是要获得 μ 和 β_j 的估计,在此基础上方可进行检验、预测等分析。为方便计算,事先对数据中所有的变量进行中心化后,在进行线性回归建模时可以略去常数项 μ。

对于参数 $\beta_j (j = 1, \cdots, p)$ 的估计,最小二乘 (least square) 是最常用的估计方法,即 $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}}\beta)^2$。利用矩阵形式表达可得 $\hat{\beta} = (x^{\mathrm{T}}x)^{-1}x^{\mathrm{T}}y$。其中:$x = (x_1, \cdots, x_n)^{\mathrm{T}}$,$x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$,$y = (y_1, \cdots, y_n)^{\mathrm{T}}$,$\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$。

众所周知,最小二乘估计是一种基于均值回归的估计方法,在误差方差并不太大或服从高斯分布时,最小二乘估计具有一些好的优点,如具有明确的估计表达式,估计量具有无偏性、相合性。然而,当收集到的数据受到干扰,即具有较大的噪声时,或者说数据的信噪比较小时,此时利用最小二乘估计进行数据挖掘效果会很差。原因在于,最小二乘估计是一种非稳健的方法,即当式(1)中的误差 ε_i 方差较大(即数据中噪声较大)或误差 ε_i 不服从高斯分布时,最小二乘估计表现很糟糕。为此,需要寻找稳健而且有效的估计方法。

1.2　特征提取(变量选择)方法

利用回归模型进行数据挖掘的特征提取时,本质上就是需要寻找出影响数据的重要变量,即变量选择问题。经典的变量选择方式有假设检验和信息准则两大类。假设检验的方法包括前进法、后退法、逐步回归法、最优回归子集法等;信息准则包括 AIC(Akaike information criterion)和 BIC(Bayesian information criterion)等。实践中,当初始模型中解释变量个数较少时,用经典方法作变量选择比较常见,但是当协变量个数较多时,就会存在以下问题:a)当解释变量个数非常多时计算量非常大;b)在逐步选取模型的过程中,会有累积的随机误差;c)在缩减的过程中缺乏稳定性[2]。

为克服以上经典变量选择的弱点,对于式(1),斯坦福大学著名统计学家 Tibshirani[3] 开创性地提出了用惩罚函数的方法同时进行变量选择和系数估计, 具体地表示为

$$\hat{\beta}^{L} = \underset{\beta}{\operatorname{argmin}} \{ \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}}\beta)^2 + n\lambda \| \beta \|_1 \} \tag{2}$$

其中:$\| \beta \|_1 = \sum_{j=1}^{p} |\beta_j|$,$\lambda$ 是一个惩罚参数,称 $\hat{\beta}^{L}$ 为正则化 LASSO(least absolute shrinkage and selection operator)估计。LASSO 估计最大的优点即它是一种连续缩减的正则化估计,能准确地选择出重要的变量,并能给出系数的估计,同时将不相关变量的系数估计为 0。随后 Efron 等人[4] 又针对于 LASSO 估计提出了有效的角回归算法(LARS),从而基于正则化 LASSO 的变量选择方法有了迅猛的发展。

需要指出的是,前面提到的角回归算法只适用于变量的维数小于或等于样本观测数目,即 $p \leqslant n$ 情形。为解决 $p > n$ 情形,Friedman 等人[5] 提出了坐标下降算法(coordinate-wise descent)的思想,其核心思想是:在极小化(或极大化)某个目标函数时,每一步只对参数的一个分量坐标进行极小化而固定参数的其他分量值不断进行循环迭代计算,直至每一个参数的分量收敛。

另外,注意到式(2)中惩罚函数对于不同的回归系数使用相同的惩罚参数 λ,这是不太合理也不太公平的。为此,Zou[6] 提出了下面的可适应 LASSO 估计方法(adaptive LASSO):

$$\hat{\beta}^{AL} = \underset{\beta}{\operatorname{argmin}} \{ \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}}\beta)^2 + n\lambda \sum_{j=1}^{p} \omega_j |\beta_j| \} \tag{3}$$

其中:ω_j 起着权重的作用,本质上就是对不同的回归系数进行不同程度的惩罚。通过设置权重 ω_j,自动对重要变量的系数进行较小的惩罚而对不相关变量的系数进行较大的惩罚,从而实现自适应的特征提取(变量选择)。

然而,可适应 LASSO 估计式(3)本质是基于最小二乘的一种惩罚估计,当数据本身噪声较大时,由 2.1 节可知最小二乘的弱点,其表现也会较差。本文将基于稳健的惩罚估计方法,并假定变量的个数远远大于样本数据的观测数目,即 $p >> n$,研究高维数据的特征提取(变量选择)问题。

2　特征提取(变量选择)的稳健方法

2.1　中位数回归

均值(mean)是数据的平均值,而中位数(median)就是把一组数据按大小顺序排列后,最中间的数据称为中位数。这两个量都是度量数据中心的重要指标。然而,均值容易受极端数据的影响,其数值会偏离主体数据的特性。中位数的优点是不受偏大或偏小数据的影响,有很好的主体数据(或称大部分数据)的特性。在线性回归模型中,为克服数据中的噪声对估计的不利影响,可以基于残差的绝对值进行回归系数的估计,称为中位数回归(median regression)[7],即

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - x_i^{\mathrm{T}}\beta| \tag{4}$$

中位数回归估计也称为最小一乘(least absolute) 估计。最小一乘估计相比于最小二乘估计最大的优点是估计受噪声大小的影响较小,具有稳健性,甚至在噪声为无穷时,也能获得不错的估计。

为清楚地看清中位数回归的原理及其稳健性,假定最简单的线性回归模型,即只有一个变量的回归模型 $y_i = \alpha + \varepsilon_i (i = 1, \cdots, n)$。很明显,参数 α 的最小二乘估计即为 $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i$,即为 y 的样本均值。然而,α 的最小一乘估计恰为 $\hat{\alpha} = \operatorname{median}(y_1, \cdots, y_n)$,即为 y 的中位数,它不受数据中部分数据噪声大的影响。当有多个变量时,可以利用式(4)而获得的估计就具有稳健性。

对于中位数估计式(4),可以采用 MATLAB 软件中线性规划程序或者直接使用 R 语言中的 QR 软件包进行计算,计算非常简洁、快速。

2.2　稳健的特征提取方法

由于中位数回归不易受数据噪声的影响,具有稳健性,可以基于此并利用正则化估计方法对高维数据进行特征提取。为提高变量选择的效果,可以采用如下的两步稳健估计方法进行特征提取。

第一步获得稳健的正则化 LASSO 估计,即

$$\hat{\beta}^{(0)} = \underset{\beta}{\operatorname{argmin}} \{ \sum_{i=1}^{n} |y_i - x_i^{\mathrm{T}}\beta| + n\lambda_1 \| \beta \|_1 \} \tag{5}$$

第二步获得稳健的正则化 adaptive LASSO 估计,即

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{ \sum_{i=1}^{n} |y_i - x_i^{\mathrm{T}}\beta| + n\lambda_2 \sum_{j=1}^{p} \omega_j |\beta_j| \} \tag{6}$$

其中：$\omega_j = 1/|\hat{\beta}_j^{(0)}|^{\gamma}$，$\gamma \geq 1$，$\hat{\beta}_j^{(0)}$ 是由第一步得到的正则化 LASSO 估计。

由 LASSO 和 adaptive LASSO 估计的稀疏性，可以通过式(5)和(6)自动得到重要的变量，同时提出不相关变量，进而实现数据的特征提取。需要强调的是，利用本文提出的方法获得的估计具有很好的性质，最重要的一点是得到的估计虽然基于全部的高维数据，但经过正则化估计后就像直接对事先知道哪些变量或数据特征是重要的进行回归估计一样的好，并且最终几乎可以非常正确地选择出重要变量和剔除不相关变量。

2.3 算法的实现

基于中位数回归的稳健正则化估计式(5)和(6)具有计算方便、快速的特点，其原因主要在于式(5)和(6)中的目标函数都是绝对值函数形式，相应的估计都可以转换成中位数回归估计来完成。具体如下：

a) 设置 $\lambda_1^{\max}$，$\lambda_2^{\max}$，令 $\lambda_1 = \frac{k}{100}\lambda_1^{\max}$，$\lambda_2 = \frac{k}{100}\lambda_2^{\max}$，$k = 1, \cdots, 100$；

b) 对每一个固定的惩罚参数 λ_1，作数据变换，令 $y^* = \begin{pmatrix} y_{n\times 1} \\ 0_{p\times 1} \end{pmatrix}$ 为 $n+p$ 维列向量，$x^* = \begin{pmatrix} x_{n\times p} \\ \lambda_1 I_{p\times p} \end{pmatrix}$ 为 $(n+p)\times p$ 矩阵[7]，利用中位数回归方法估计。

$$\hat{\beta}^{(0)}(\lambda_1) = \underset{\beta}{\operatorname{argmin}} \| y^* - x^*\beta \|_1$$

c) 首先选择出最优的惩罚参数 λ_1^{opt}，即

$$\lambda_1^{\text{opt}} = \underset{\lambda_1}{\operatorname{argmin}} \{ \| y - x\hat{\beta}^{(0)}(\lambda_1) \| + \log(n) \cdot df_1 \}$$

其中：df_1 为 $\hat{\beta}^{(0)}(\lambda_1)$ 中不为 0 的分量个数。然后，获得估计 $\hat{\beta}^{(0)}$，即

$$\hat{\beta}^{(0)} = \hat{\beta}^{(0)}(\lambda_1^{\text{opt}})$$

d) 令 $\omega_j = \min\{1/|\hat{\beta}^{(0)}|, 10^6\}$，$j = 1, \cdots, p$；$W = \operatorname{diag}(\omega_1, \cdots, \omega_p)$；

e) $x^{**} = \begin{pmatrix} x_{n\times p} \\ \lambda_2 W_{p\times p} \end{pmatrix}$ 为 $(n+p)\times p$ 矩阵，利用中位数回归方法估计。

$$\hat{\beta}(\lambda_2) = \underset{\beta}{\operatorname{argmin}} \| y^* - x^{**}\beta \|_1$$

f) 选择出最优的惩罚参数 λ_2^{opt}，即

$$\lambda_2^{\text{opt}} = \underset{\lambda_2}{\operatorname{argmin}} \{ \| y - x\hat{\beta}(\lambda_2) \| + \log(n) \cdot df_2 \}$$

其中：df_2 为 $\hat{\beta}(\lambda_2)$ 中不为 0 的分量个数。获得最终估计 $\hat{\beta} = \hat{\beta}(\lambda_2^{\text{opt}})$。

3 仿真模拟实验结果

本章通过两个模拟实验例子，利用本文提出的方法进行特征提取，并将其结果与最小二乘的结果进行比较来验证本文提出方法的有效性。

3.1 模拟实例 1

模拟数据由以下方式生成：

$$y = x^{\mathrm{T}}\beta + \varepsilon$$

其中：$\beta = (3, 1.5, 2, 0, 0, 0, 0, 0)^{\mathrm{T}}$，相当于在此数据中总共 8 个变量，其中 3 个是重要变量，另外 5 个变量是不相关变量，自变量 x 服从多元正态分布 $x \sim N(0, \Sigma)$，协方差阵 $\Sigma_{ij} = 0.8^{|i-j|}$ $(1 \leq i, j \leq 8)$。ε 是误差噪声，设置四种不同的误差噪声来源，即：a) 高斯分布 $N(0,1)$，方差为 1；b) 自由度为 3 的 $t(3)$ 分布，方差为 3；c) 混合高斯分布 $0.9N(0,1) + 0.1N(0,10)$，方差为 10.9；d) 柯西分布，方差为无穷大。在不同的误差噪声下，其相应的信噪比($\operatorname{var}(y)/\operatorname{var}(\varepsilon)$)分别为 35.93、12.6433、4.2046 和 1。在模拟中，每次生成两个样本 0 数据集，即训练数据集和检验数据集，其样本量都为 60。训练数据集用于估计和选择重要变量，检验数据集用于评价估计的预测偏差。对于每一种不同误差噪声，重复模拟 200 次。

对每次生成的数据利用本文提出的算法进行估计和变量选择，并将变量选择的结果与其真值进行比较，主要包括正确选择重要变量的个数、错误选择不重要变量的个数(200 次的平均值)、用检验数据集预测的均方误差。为了比较效果，也同时报告出最小二乘正则化估计方法[7]得到的结果。模拟结果如表 1 所示。

表 1 实例 1 模拟分析结果

噪声来源	方法	正确选择不重要变量数目	错误选择重要变量数目	预测偏差
高斯分布	中位数回归	4.995 0	0.025 0	0.243 5
	最小二乘	4.990 0	0.010 0	0.174 9
t 分布	中位数回归	4.990 0	0.055 0	0.297 4
	最小二乘	4.615 0	0.070 0	0.345 5
混合高斯	中位数回归	4.995 0	0.035 0	0.237 8
	最小二乘	4.770 0	0.045 0	0.285 0
柯西分布	中位数回归	4.885 0	0.255 0	0.528 8
	最小二乘	1.280 0	0.310 0	9.867 2
理想真值		5	0	——

3.2 模拟实例 2

数据的生成方式与实例 1 基本相似，但 β 是一个 200 维变量，其前三个分量取值为 3，1.5，2，其余分量全为 0，自变量 x 仍然服从多元正态分布 $x \sim N(0, \Sigma)$，协方差阵 $\Sigma_{ij} = 0.8^{|i-j|}$ $(1 \leq i, j \leq 200)$。此数据中，总共有 200 个变量，其中 3 个是重要变量，其余 197 个变量为不相关的噪声变量。模拟结果如表 2 所示。

表 2 实例 2 模拟分析结果

噪声来源	方法	正确选择不重要变量数目	错误选择重要变量数目	预测偏差
高斯分布	中位数回归	197.000 0	0.045 0	0.262 3
	最小二乘	196.820 0	0.010 0	0.211 0
t 分布	中位数回归	196.940 0	0.055 0	0.305 5
	最小二乘	191.985 0	0.050 0	0.704 0
混合高斯	中位数回归	196.900 0	0.085 0	0.446 7
	最小二乘	171.800 0	0.130 00	1.286 1
柯西分布	中位数回归	187.400 0	0.510 0	3.355 3
	最小二乘	158.600 0	0.830 0	11.111 2
理想真值		197	0	——

从表 1 和 2 清楚地看出，基于中位数回归的正则化估计整体上优于最小二乘基于的正则化估计效果。随着信噪比的降低，其优势越来越明显，特别当噪声来源是柯西分布时，最小二乘基于的正则化估计效果特别差，而利用本文提出的方法效果仍然不错，具有很好的稳健性。另外，在第二个模拟实例中，虽然训练样本数据集中变量数目(200)远大于样本数目(60)，但本文提出的稳健的正则化估计方法效果依然很好。

4 结束语

高维数据挖掘的特点是数据中变量个数 (下转第 413 页)

表 3　基于任务量的维修专业关联度

TID	S_1	S_2	S_3	S_4	S_5	S_6	MSG
T_1	0.67	0.25	1	0.75	0.58	–	0.6
T_2	0.91	1	–	0.91	0.6	0.45	0.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
MIS	0.5	0.5	0.5	0.5	0.5	0.5	0.8

然后,根据各维修专业在维修任务集中出现的频度,可以得到其相应的支持度以及满足最小支持度的频繁项集 L_1,如表 4 所示。

表 4　专业频繁项集 L_1(正式)

Itemset	Support	Itemset	Support
S_1	0.65	S_4	0.6
S_3	0.5	S_5	0.5

最后,可以得到满足最小关联度和最小支持度的维修专业组合方案即频繁集,如表 5 所示。

表 5　关联规则输出表

关联规则	支持度	置信度	关联规则	支持度	置信度
$S_4 \rightarrow S_1$	0.45	0.75	$S_4 \rightarrow S_3$	0.45	0.75

其中,关联规则 $S_4 \rightarrow S_1$ 的含义是 45% 的维修任务同时需要光学专业 S_3 和电子专业 S_4,75% 的维修任务需要电子专业 S_4 的同时还会需要光学专业 S_3,关联规则 $S_4 \rightarrow S_3$ 也具有同样的意义。此外,根据表 2 中可以累计得到各维修专业的任务量,如表 6 所示。

表 6　维修专业任务量累计表　　/h

	机械专业 S_1	通信专业 S_2	光学专业 S_3	电子专业 S_4	液压专业 S_5	计算机专业 S_6
任务量	92	182	83	87	204	189.5

由此可见,维修专业 S_4 和 S_3 组合后其任务量为 170;维修专业 S_4 和 S_1 组合后其任务量为 179,与其他维修专业的任务量基本相当。从分析结果来看,电子专业 S_4 既可以与机械专业 S_1 进行组合,也可以与光学专业 S_3 进行组合。显然,专业组合方案 $\{S_1, S_4\}$ 和 $\{S_3, S_4\}$ 在任务共担方面存在冲突,根据式(2)和(3),其评估结果如表 7 所示。

表 7　基于支持度和置信度的评估结果

	$S_4 \rightarrow S_1$	$S_4 \rightarrow S_3$
皮尔森系数 α	0.25	0.61
余弦函数 β	0.72	0.82

可见,随着维修任务的增加,光学专业与电子专业共同承担任务的趋势或几率更大。由此可见,与机械专业 S_1 相比,光学专业 S_3 与电子专业 S_4 的组合是最优的。

5　结束语

传统的数据统计方法难以发现维修专业隐藏在任务信息中潜在的关联关系,而基于数据挖掘的维修专业组合优化方法是切实可行的。通过引入维修专业关联度,利用改进的多最小支持度关联规则挖掘任务信息中维修专业潜在的规律,既实现了任务共担的维修专业有机组合,也均衡了各维修专业的任务量,提高了组合专业协同维修的效率和综合保障能力。

参考文献:

[1] 赵卫民,吴勋,孟宪君,等. 武器装备论证学[M]. 北京:兵器工业出版社,2008:31-32.

[2] 毛国君,段立娟,王实,等. 数据挖掘原理与算法[M]. 北京:清华大学出版社,2007:67-68.

[3] MABU S, SHIMADA K. Fuzzy inter transaction class association rule mining using genetic network programming for stock market prediction [J]. IEEE Trans on Electrical and Electronic Engineering,2011, 6(4):2697-2750.

[4] ZHANG Xiao-hui, LIU Zheng-jiang. Analysis of multi-dimension association rule in marine casualties[C]//Proc of the 1st International Conference on Transportation Information and Safety. 2011: 2697-2705.

[5] 刘晶,季海鹏,朱清香. 改进多重最小支持度关联规则算法在故障诊断中的应用[J]. 工业工程,2010,13(4):108-111.

[6] ENGLE K M, RADA R. A top-k analysis using multilevel association rule mining for autism treatments[C]//Proc of the 6th International Conference on Universal Access in Human-Computer Interaction. Berlin: Springer-Verlag,2011:328-334.

[7] 周永生,熊结青,沙宗尧. 基于关联规则挖掘的生化企业数据分析及其应用研究[J]. 计算机与应用化学,2010,27(9):1252-1256.

[8] 叶义成,柯丽华,黄德育,等. 系统综合评价技术及其应用[M]. 北京:冶金工业出版社,2006:18-19.

[9] 迈克伦南,唐朝晖,克里沃茨. 数据挖掘原理与应用[M]. 董艳,程文俊,译. 北京:清华大学出版社,2007.

[10] 何朝阳,赵剑锋. 最大控制的多最小支持度关联规则及其挖掘方法研究[J]. 计算机工程,2006,32(11):103-105.

[11] 徐建中,朱建新. 应用统计学[M]. 哈尔滨:哈尔滨工程大出版社. 2001:283-284.

[12] 孙文爽,陈兰祥. 多元统计分析[M]. 北京:高等教育出版社. 1999:341-343.

(上接第 376 页)往往大于样本观测数目,通常数据中还存在较大的噪声。本文在中位数回归分析方法的基础上提出了一种新的有效的变量选择(或特征提取)方法,并具体给出了估计算法,该算法具有快速计算的特点。从实验结果可以看出,该方法能更加有效地进行变量选择(特征提取),且预测偏差较小,达到了良好的效果。当然,在实际问题中,数据的类型具有多样性、复杂性,噪声的来源也是多渠道的,采用一种方法并不能有效地解决所有的问题。例如,当高维数据中变量是属性数据或离散的计数数据,直接利用本文提出的方法进行分析,效果并不理想。关于这方面的问题还将作进一步研究。

参考文献:

[1] HASTIE T, TIBSHIRANT R, FRIEDMAN J. 统计学习基础:数据挖掘、推理与预测[M]. 范明,等译. 北京:电子工业出版社,2004.

[2] BüHLMANN P, Van de GEER S. Statistics for high-dimensional data: methods, theory and applications[M]. Berlin: Springer-Verlag, 2011.

[3] TIBSHIRANI R. Regression shrinkage and selection via the LASSO [J]. Journal of the Royal Statistical Society: Series B,1996,58 (1): 267-288.

[4] EFRON B, HASTIE T, JOHNSTONE I, *et al*. Least angle regression [J]. The Annals of Statistics,2004,32(2): 407-489.

[5] FRIEDMAN J, HATTIE T, HOFLING T, *et al*. Pathwise coordinate optimization[J]. Annals of Applied Statistics,2007,1(2):302-332.

[6] ZOU Hui. The adaptive LASSO and its oracle properties[J]. Journal of the American Statistical Association,2006,101(476): 1418-1429.

[7] WANG Han-sheng, LI Guo-dong, JIANG Guo-hua. Robust regression rhrinkage and consistent variable selection through the LAD-Lasso[J]. Journal of Business and Economic Statistics,2007,25 (3): 347-355.