

高维数据挖掘中特征选择的稳健方法

李泽安^{1*}, 陈建平¹, 章雅娟¹, 赵为华²

(1. 南通大学 计算机科学与技术学院, 江苏 南通 226019; 2. 南通大学 理学院, 江苏 南通 226019)

(* 通信作者电子邮箱 lizean@ntu2012@163.com)

摘要: 针对高维数据的特点, 即数据中变量个数往往大于样本观测数目, 并且数据往往具有异质性特点, 基于众数回归分析和变量选择降维技术, 提出了一种稳健有效的特征选择方法, 利用局部二次逼近算法 (LQA) 和最大期望 (EM) 算法, 给出估计算法和最优调节参数的选取方法。通过实验的模拟数据分析表明, 所提出的特征提取选择方法整体优于基于最小二乘和中位数的正则化估计方法, 特别当误差是非正态分布时, 与已有方法相比具有较高的预测能力和稳健性。

关键词: 高维数据; 特征选择; 众数回归; 自适应 LASSO; 最大期望算法

中图分类号: TP311 **文献标志码:** A

Robust feature selection method in high-dimensional data mining

LI Zean^{1*}, CHEN Jianping¹, ZHANG Yajuan¹, ZHAO Weihua²

(1. College of Computer Science and Technology, Nantong University, Nantong Jiangsu 226019, China;

2. College of Science, Nantong University, Nantong Jiangsu 226019, China)

Abstract: According to the feature of high-dimensional data, the number of variables is usually larger than the sample size and the data are often heterogeneous, a robust and effective feature selection method was proposed by using the dimensional reduction technique of variable selection and the modal regression based estimation method. The estimation algorithm was given by using Local Quadratic Algorithm (LQA) and Expectation-Maximum (EM) algorithm, and the selection method of the parameter adjustment was also discussed. Data analysis of the simulation shows that the proposed method is overall better than the least square and median regression based regularized method. Compared with the existing methods, the proposed method has higher prediction ability and stronger robustness especially for the non-normal error distribution.

Key words: high-dimensional data; feature selection; modal regression; adaptive Least Absolute Shrinkage and Selection Operator (LASSO); Expectation-Maximum (EM) algorithm

0 引言

为对实际问题中收集到的数据进行有效的特征选择, 尽可能地挖掘出数据中潜在的、有用的信息, 需要对数据事先进行一些数据分析。数据分析的目的是从隐藏在一大批看来杂乱无章的数据中找出所研究对象的内在规律。

文献[1]提出, 典型的数据分析包含以下步骤: 1) 探索性数据分析。对从实际问题中收集到的数据, 通过作图 (如直方图、箱线图和 QQ 图)、造表 (如数据分类表和属性分析表)、用各种形式的曲线拟合 (如线性曲线、非线性曲线等) 或计算某些特征量等手段探索规律性的可能形式。2) 选择恰当的模型。提出一类或几类可能的统计模型或数学模型, 通过进一步的分析从中挑选一定的模型, 当然最终挑选哪一种模型需要综合各方面进行评价分析。3) 统计推断分析。通常使用统计学方法对所选定的模型进行相关的统计推断 (如参数估计、区间估计和假设检验等), 根据推断的结果给出合理的解释和分析, 进而实现从数据中提取出有用的信息。

然而, 从实际问题中收集得到的数据含有变量的个数远远大于数据的观测数目, 称为高维数据。高维数据挖掘与传统的数据挖掘相比较最主要的特点在于它的维度 (属性) 通

常可以达到成百上千维, 甚至更高。

许多高维数据的一个共同特征是具有变量的稀疏性。利用稀疏性特征, 可以从成百上千维变量中有效地选择出真实的影响变量, 从而达到特征选择的目的, 这是进行高维数据挖掘需要解决的问题。对于高维数据常用做法是通过降维将数据从高维降低到低维, 然后用低维数据的处理办法进行处理。另一方面, 高维数据具有“异质性”特点, 即数据中含有异常值, 数据中具有较大的噪声, 即信噪比不高以及数据之间一般不独立具有较强的相关性, 使得对高维数据进行有效的特征选择具有很大的挑战性。

本文在探索性数据分析的基础上, 利用统计学中的回归分析方法提出高维数据特征选择的有效方法。该方法能克服已有方法的弱点, 特别是在对异常点、异方差和相关性较强的数据进行特征提取时具有很好的稳健性, 能有效地进行特征提取, 真正挖掘出隐藏在数据内部的有用信息。

1 探索性数据分析

1.1 一维探索性数据分析

在数据挖掘的一维探索性分析^[1]中, 均值 (mean)、中位

收稿日期: 2013-03-11; 修回日期: 2013-05-06。

基金项目: 南通大学杏林学院自然科学基金资助项目 (2012K116); 南通大学自然科学基金资助项目 (11Z067)。

作者简介: 李泽安 (1977 -), 女, 江苏南通人, 讲师, 硕士, CCF 会员, 主要研究方向: 数据挖掘; 陈建平 (1960 -), 男, 江苏南通人, 教授, 主要研究方向: 数据分析; 章雅娟 (1977 -), 女, 甘肃白银人, 讲师, 硕士, 主要研究方向: 数据挖掘; 赵为华 (1978 -), 男, 江苏海门人, 讲师, 博士, 主要研究方向: 统计学。

数(median)和众数(mode)是度量数据“中心”的三个最重要的数字特征。均值即为观测数据的平均值,由于均值具有理解直观、计算方便等特点,在实际中经常被使用;数据的中位数值是指将观测数据从小到大排序后最中间的那个数为数据的中位数;而数据的众数值是指数据中取值频率最高的那个值。

均值虽然简单直观,但它容易受样本数据中异常值的影响^[2-3],其取值有时可能偏离数据主体,此时均值就不能反映出数据真实的特点。譬如,一个公司有100位员工,其中10位管理人员每人年薪100万,其余90人是普通员工每人年薪大约3万,公司在招聘新人时宣称员工的平均年薪为12.7万,显然该公司的平均年薪就不能很好反映员工工资的特点。但中位数和众数就不受少数异常值的影响,仍能通过数据反映事物本质的特征。相比而言,众数更能集中反映数据主体的取值特点,它反映了数据中最有可能(most likely)的取值,因而能够提供更丰富的信息。

1.2 多维数据探索性分析

为进一步探索数据中因变量(响应变量)与自变量(解释变量)之间的关系,即数据的多维数据探索性分析,回归分析是行之有效的办法之一。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。然后,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,特征选择(或称变量选择)是数据有效数据挖掘的一个重要步骤。

经典的线性回归模型可表示为

$$y_i = \mu + \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i; i = 1, 2, \dots, n \quad (1)$$

其中: y_i 是因变量; x_{ij} 是解释变量; μ 是常数项; $\beta_j (j = 1, 2, \dots, p)$ 是回归系数; ε_i 是模型误差,其均值为0,方差为 σ^2 。不失一般性,可假设 $\mu = 0$, 否则可以对响应变量中心化即可。对于模型(1),首要问题是要获得 $\beta_j (j = 1, 2, \dots, p)$ 的估计,在此基础上方可进行检验分析、预测分析等。对于参数 β_j 的估计,最小二乘(Least Square)是最常用的估计方法,即

$$\hat{\beta}^{LS} = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i^T \beta)^2 \quad (2)$$

最小二乘估计是一种基于均值回归的估计方法,在误差方差不太大或服从高斯分布时,最小二乘估计有明确的估计表达式,估计量具有无偏性和相合性。然而,当收集到的数据受到干扰,即数据异质性特点时,此时利用最小二乘估计进行数据挖掘不具有稳健性,效果会很差。

为解决均值回归的弱点,研究者寻找其他的方法进行系数估计,最常用的是中位数回归^[4](median regression)或称最小一乘(least absolute)方法,即

$$\hat{\beta}^{LA} = \arg \min_{\beta} \sum_{i=1}^n |y_i - x_i^T \beta| \quad (3)$$

然而,中位数回归虽然具有稳健性特点,但是中位数回归估计中含有绝对值,在数学上处理不方便。另外,当数据服从正态分布或数据不含有异常值时,中位数回归估计的效率会降低。为此,本文将基于众数回归(modal regression)方法研究系数的估计,该方法借鉴非参数估计的思想,得到的估计不

仅具有稳健性而且还是一个非常有效的估计。进一步,当数据是高维时,提出基于众数回归的惩罚估计方法或称正则化估计方法,进而实现对数据的特征选择。

2 特征提取的稳健方法

2.1 众数回归

关于众数回归估计已有一些文献进行研究:如文献[5]中提出使用均匀核和固定的窗宽研究回归系数估计;文献[6]中,在研究多元密度函数估计方法时,提及了众数回归并解释了众数回归的优点,但并没有给出如何实施估计和统计推断的具体方法;最近,文献[7]中基于众数回归思想系统地研究了非参数回归的稳健估计,并给出了估计的理论性质,讨论了估计的实施细节。

对于模型(1),借鉴文献[7]众数回归估计的思想,系数 β 的估计定义为

$$\hat{\beta}^{MR} = \arg \max_{\beta} \frac{1}{n} \sum_{i=1}^n \varphi_h(y_i - x_i^T \beta) \quad (4)$$

其中: $\varphi_h(t) = \frac{1}{h} \varphi(t/h)$, $\varphi(\cdot)$ 是核函数(kernel function); h 称为窗宽(bandwidth),窗宽 h 在估计中起着重要的作用,通过调节 h 的取值使得估计具有很好的稳健性,具体的选择方法将在下一节给出。称由式(4)得到的估计为众数回归估计。

为展示众数回归估计的原理及其稳健性,假定最简单的线性回归模型,即只有一个变量的回归模型 $y_i = \alpha + \varepsilon_i (i = 1, 2, \dots, n)$ 。显然参数 α 的最小二乘估计和最小一乘估计分别为 $\hat{\alpha}^{LS} = \frac{1}{n} \sum_{i=1}^n y_i$ 和 $\hat{\alpha}^{LA} = \text{median}(y_1, y_2, \dots, y_n)$ 。然而 α 的众

数回归估计为 $\hat{\alpha}^{MR} = \arg \max_{\alpha} \frac{1}{n} \sum_{i=1}^n \varphi_h(y_i - \alpha)$ 。事实上,由于 $\frac{1}{n} \sum_{i=1}^n \varphi_h(y_i - \alpha)$ 为误差 ε 的密度函数估计,对此求极大值就可以获得使密度函数达到最大值时的估计值,因此称 $\hat{\alpha}^{MR}$ 为众数估计,它反映了最可能的取值估计,此估计不受数据中噪声大而具有异质性的影响,因而获得的估计就具有稳健性。同时,基于众数估计进一步可以获得系数的区间估计,所得的区间估计长度要比其他估计短,即推断精度高,能提供更有效的信息。

2.2 基于众数回归的惩罚估计

利用回归模型对数据进行特征选择时,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。斯坦福大学著名统计学家 Tibshirani 在文献[8]中基于最小二乘估计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

$$\hat{\beta}^{PLS} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - x_i^T \beta)^2 + n\lambda \|\beta\|_1 \right\} \quad (5)$$

其中: $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, λ 是一个惩罚参数。称 $\hat{\beta}^{PLS}$ 为正则化 LASSO(Least Absolute Shrinkage and Selection Operator)估计。LASSO 估计最大的优点在于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。随后文献[9]中又针对 LASSO 估计提出了有效的角回归算法(Least Angle Regression, LARS),从而基于正则化 LASSO 的变量选择方法有了迅猛的发展。

然而,基于最小二乘惩罚估计方法的弱点与无惩罚的最

小二乘估计一样,不是一个稳健估计,受异常点的影响很大。为此,提出基于众数回归的惩罚估计如下:

$$\hat{\beta}^{\text{PMR}} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \varphi_h(y_i - \mathbf{x}_i^T \beta) - n\lambda \|\beta\|_1 \right\} \quad (6)$$

注意到式(4)中惩罚函数对于不同的回归系数使用相同的惩罚参数,这既不太合理也不太公平。为此,借鉴文献[10]的自适应 LASSO 估计思想,提出如下的自适应惩罚估计方法:

$$\hat{\beta}^{\text{APMR}} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \varphi_h(y_i - \mathbf{x}_i^T \beta) - n\lambda \sum_{j=1}^p \omega_j |\beta_j| \right\} \quad (7)$$

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选择。

3 算法实现

3.1 估计算法

为实现稳健的自适应特征提取,提出如下的两步估计方法进行特征提取:

第一步获得正则化 LASSO 型估计,即

$$\hat{\beta}^{\text{PMR}} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \varphi_h(y_i - \mathbf{x}_i^T \beta) - n\lambda \|\beta\|_1 \right\} \quad (8)$$

第二步获得正则化 Adaptive LASSO 型估计,即

$$\hat{\beta}^{\text{APMR}} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \varphi_h(y_i - \mathbf{x}_i^T \beta) - n\lambda \sum_{j=1}^p \omega_j |\beta_j| \right\} \quad (9)$$

其中: $\omega_j = \frac{1}{|\hat{\beta}_{\text{PMR}}^j|^\gamma}$, $\gamma \geq 1$, $\hat{\beta}_{\text{PMR}}^j$ 是由第一步得到的稳健正则化 LASSO 估计。

但是,直接极大化式(8)或式(9)有时非常困难,一方面由于惩罚函数在 0 处不可导,另外一方面是整个目标函数不是凸函数。下面将基于局部二次逼近算法(Local Quadratic Algorithm, LQA)^[11]以及最大期望(Expectation-Maximum, EM)算法^[12]提出易于实施的算法。下面叙述是基于正则化 Adaptive LASSO 估计式(9),式(8)中的估计获得类似。由于式(4)中的核函数 $\varphi(\cdot)$ 选取对最终的估计影响不大,为方便起见,取 $\varphi(\cdot)$ 为标准正态密度函数核。

由于绝对值函数在 0 处不可导,应用 LQA, $|\beta_j| \approx \sqrt{\beta_j^2 + c}$, 其中 c 是一个非常小的正数。令 $m = 0$ 和 $\beta_j^{(m)}$ 为第 m 步迭代时的估计值,记

$$\Sigma^{(m)} = \text{diag} \left(\frac{\omega_1 \beta_1^{(m)}}{\sqrt{(\beta_1^{(m)})^2 + c}}, \frac{\omega_2 \beta_2^{(m)}}{\sqrt{(\beta_2^{(m)})^2 + c}}, \dots, \frac{\omega_p \beta_p^{(m)}}{\sqrt{(\beta_p^{(m)})^2 + c}} \right); \beta^{(m)} = (\beta_1^{(m)}, \beta_2^{(m)}, \dots, \beta_p^{(m)})^T$$

进一步应用 EM 算法,提出如下算法:

步骤 1(E-step): 计算

$$\pi(i | \beta^{(m)}) = \frac{\varphi_h(y_i - \mathbf{x}_i^T \beta^{(m)})}{\sum_{j=1}^n \varphi_h(y_j - \mathbf{x}_j^T \beta^{(m)})}; i = 1, 2, \dots, n$$

步骤 2(M-step): 令 $W = \text{diag}(\pi(1 | \beta^{(m)}), \pi(2 | \beta^{(m)}), \dots, \pi(n | \beta^{(m)}))$, 计算

$$\beta^{(m+1)} \approx \arg \max_{\beta} \left(\sum_{i=1}^n \pi(i | \beta^{(m)}) \log \varphi_h(y_i - \mathbf{x}_i^T \beta) - \right.$$

$$n\lambda \sum_{j=1}^p \omega_j \sqrt{(\beta_j^{(m)})^2 + c} \Big) \approx (X^T W X + n\lambda \Sigma^{(m)})^{-1} X^T W Y$$

其中: $X = (x_1, x_2, \dots, x_n)^T$, $Y = (y_1, y_2, \dots, y_n)^T$ 。

步骤 3: 令 $m = m + 1$, 重复迭代步骤 1 和步骤 2 直至收敛。

记 $\hat{\beta}^{\text{APMR}}$ 为 $\beta^{(m)}$ 的最终迭代值。

由 Adaptive LASSO 估计的稀疏性,可以通过以上算法自动得到重要的变量,同时剔除不相关变量^[13-15],进而实现数据中的特征提取。利用本文提出的方法获得的估计具有很好的性质,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有效,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,而且估计具有很好的稳健性。

3.2 窗宽和惩罚参数选择

实施上一节提出的算法,还需要选取合适的窗宽和惩罚参数。建议使用下面的方法选择最优的窗宽。

$$\text{令 } \hat{F}(h) = \frac{1}{n} \sum_{i=1}^n \ddot{\varphi}_h(\hat{\varepsilon}_i), \hat{G}(h) = \frac{1}{n} \sum_{i=1}^n (\dot{\varphi}_h(\hat{\varepsilon}_i))^2, \text{ 其}$$

中 $\hat{\varepsilon}_i = y_i - \mathbf{x}_i^T \hat{\beta}^{\text{int}}, \hat{\beta}^{\text{int}}$ 为初始估计,可以利用基于最小二乘的惩罚估计获得。令 $R(h) = \hat{G}(h) (\hat{F}(h))^{-2}$, 然后通过格子点方法选取最优的窗宽 $h_{\text{opt}} = \arg \min_h R(h)$, 比较好的格子点方法可以这样设定: $h = 0.5 \cdot \hat{\sigma} \cdot 1.02^j$ ($j = 1, 2, \dots, 100$), 这

$$\text{里 } \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\varepsilon}_i)^2}.$$

关于惩罚参数 λ 的选取,可以通过 BIC (Bayesian Information Criterion) 标准进行: $\lambda_{\text{opt}} = \arg \max_{\lambda} BIC(\lambda)$, 其中

$$BIC(\lambda) = \sum_{i=1}^n \varphi_h(y_i - \mathbf{x}_i^T \beta(\lambda)) - df_{\beta(\lambda)} \cdot n \log n, df_{\beta(\lambda)} \text{ 为估计中非 0 系数的个数。}$$

4 仿真模拟实验分析

下面给出两个模拟实验来说明本文方法的效果。为展示该方法在特征提取方面的效果,将其结果与最小二乘以及作者在文献[4]中提出的中位数估计方法的结果进行比较。

4.1 模拟实验一

模拟数据由以下方式生成:

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i; i = 1, 2, \dots, n$$

其中: $\beta = (\beta_1, \beta_2, \dots, \beta_{10})^T$ 是一个 10 维参数向量,其前三个分量取值为 2, 1, 2, 其余分量取值为 0; 自变量 x_i 服从多元正态分布 $x_i \sim N(0, \Sigma)$, 协方差阵 $\Sigma_{jk} = 0.5^{|j-k|}$ ($1 \leq j, k \leq 10$); ε_i 是误差噪声且与自变量独立,考虑三种不同的误差噪声来源,包括高斯分布 $N(0, 1)$ 、自由度为 3 的 $t(3)$ 分布和混合高斯分布 $0.9N(0, 1) + 0.1N(0, 10)$ 。在此数据中,总共有 10 个变量,其中 3 个是重要变量,另外 7 个变量是不相关的冗余变量。

在模拟中,每次生成两个样本数据集,即训练样本数据集和检验样本数据集,其样本量分别为 60 和 100。训练数据集用于估计和选择重要变量,检验样本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。采用本文算法对每次生成的数据进行估计和特征提取,同时将结果与其真值比较,比较主要包括以下几个方面:1) 正确识别不重要变量的个数;2) 正确选择重要变量的个数(200 次的平均值);3) 完全正确选择模型的比例;4) 系数估计与真值的绝对值偏差之和;5) 使用检验数据集得到的预测均方误差。同时采用最小二乘正则化估计方法和文献[4]中的基于中位数回归进行模拟,得到的结果如表 1 所示。

表1 实验一的模拟结果

误差来源	方法	正确识别 不重要变量个数	正确选择 重要变量个数	正确选择 模型的比例	绝对值偏差和	预测均方误差
正态分布	最小二乘估计	6.985	2.990	0.975	0.4196	1.0832
	中位数估计	6.980	2.990	0.970	0.4612	1.0920
	众数回归估计	6.985	2.995	0.980	0.4214	1.0860
$t(3)$ 分布	最小二乘估计	6.940	2.930	0.925	0.5620	3.2552
	中位数估计	6.950	2.945	0.935	0.5411	3.1502
	众数回归估计	6.965	2.980	0.960	0.5372	3.1133
混合分布	最小二乘估计	6.890	2.850	0.840	0.7820	7.7258
	中位数估计	6.905	2.905	0.885	0.5610	5.9065
	众数回归估计	6.940	2.965	0.935	0.4273	5.2969

4.2 模拟实验二

数据由以下模型生成:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + (1 + |x_{i1}|) \varepsilon_i; i = 1, 2, \dots, n,$$

其中 $\boldsymbol{\beta}$ 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 \mathbf{x}_i 服从多元正态分布 $\mathbf{x}_i \sim N(0, \boldsymbol{\Sigma})$,

协方差阵 $\boldsymbol{\Sigma}_{jk} = 0.8^{|j-k|}$ ($1 \leq j, k \leq 200$); ε_i 的选取与实验一相同。训练数据样本量和检验数据样本量分别为 100 和 200。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,并且是一个异方差的高维数据模型。模拟结果见表 2。

表2 实验二的模拟结果

误差来源	方法	正确识别 不重要变量个数	正确选择 重要变量个数	正确选择 模型的比例	绝对值偏差和	预测均方误差
正态分布	最小二乘估计	193.925	2.965	0.920	0.9290	1.2116
	中位数估计	193.825	2.955	0.910	0.9320	1.1820
	众数回归估计	194.950	2.980	0.945	0.8423	1.1220
$t(3)$ 分布	最小二乘估计	190.250	2.955	0.885	1.2305	2.7430
	中位数估计	192.200	2.950	0.915	1.0526	2.5802
	众数回归估计	192.870	2.975	0.920	1.0230	2.5225
混合分布	最小二乘估计	187.125	2.910	0.850	1.6420	10.0230
	中位数估计	189.540	2.935	0.890	1.3452	9.0235
	众数回归估计	191.550	2.955	0.915	1.2820	8.8420

观测表 1 和 2 中的实验结果,显然本文提出的特征选择方法整体上明显优于最小二乘和基于中位数的正则化估计方法,特别当误差是非正态分布时,本文提出的基于众数回归的特征提取方法具有很好的稳健性。即使在正态误差分布情形,本文方法几乎与最小二乘的正则化估计效果相当,同时要优于基于中位数的方法。值得注意的是在第二个实验中,由于是异方差模型,即使在正态误差情形,本文方法的效果甚至还比最小二乘方法略好一些,而当误差是混合正态时,本文方法的优势会更加明显。

4.3 模拟实验的时间复杂度分析

利用 EM 算法和 LQA 的思想,本文提出的具有稳健性的特征选择方法主要是基于自适应的非参数估计思想的众数回归估计。事实上,不使用 EM 算法,直接应用 LQA 直接极大化式(7),也可以得到估计,实现最终的特征选择目的。而本文使用 EM 算法起到了加速计算的功能,即如果得到了第一步

$$\pi(i|\boldsymbol{\beta}^{(m)}) = \frac{\varphi_h(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(m)})}{\sum_{j=1}^n \varphi_h(y_j - \mathbf{x}_j^T \boldsymbol{\beta}^{(m)})} (i = 1, 2, \dots, n) \text{ 的值,第}$$

二步就相当于基于加权版本最小二乘的特征选择方法。因此,新算法增加的复杂度主要在于第一步中需要计算每一个 $\varphi_h(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(m)})$,而计算 $\varphi_h(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(m)})$ 可直接调用 Matlab 中正态密度核函数的命令,也就是使用 EM 算法后将非线性的优化问题转化为常规的加权版本最小二乘的计算问题。因此,在理论上本文没有考虑算法的时间复杂度问题。另外,根据实践经验,上述估计一般只需迭代计算几步就能得到结果,

所用的计算时间与原有方法相比增加得并不多^[14]。

5 结语

本文基于众数回归估计思想并结合正则化估计方法研究了高维数据的特征选择方法,并详细给出了估计算法的实施细节以及窗宽和惩罚参数的选取方法,该算法具有快速计算的特点。实验结果证明本文提出的方法能更加有效地进行特征提取,且与已有方法相比具有较小的偏差和预测均方误差,所有结果均显示新方法具有很好的稳健性和有效性。

本文提出的特征选择的稳健方法是在线性模型框架下研究的,而在实际问题中数据和噪声具有多样性、复杂性的特点,线性模型未必适合实际的数据,因此需要寻找其他更合适的模型,如非线性模型、非参数模型和半参数模型等来研究高维数据的特征提取方法。本文提出的方法主要针对于因变量是连续型变量情形,当因变量是离散型计数数据或属性效应变量时,本文的方法不理想或不适用,需要研究更好的稳健特征提取方法。

参考文献:

- [1] GIUDICI P. 实用数据挖掘[M]. 袁方,王煜,王丽娟,等译. 北京:电子工业出版社,2004:120-140.
- [2] HAN J, KAMBER M. 数据挖掘概念与技术[M]. 范明,译. 北京:机械工业出版社,2001:98-128.
- [3] HASTIE T, TIBSHIRANT R, FRIEDMAN J. 统计学习基础:数据挖掘、推理与预测[M]. 范明,译. 北京:电子工业出版社,2004:15-70.

求发送相关授权文件,发送端通过认证提取端是否合法决定是否发送授权文件。在得到授权文件后,首先使用可逆水印算法提取出含水印图像和加密压缩补偿向量,接着再按照前文所述的提取算法通过补偿向量提取出水印和原始图像。由于补偿码是经过安全处理的,提取端如果强行恢复,则无法获得正确有意义的原始图像。所以在没有授权文件的情况下,提取端用户很难获取正确的原始图像。图5(a)和(b)分别显示了从使用控制因子为0.19得到的发布图像正确恢复和强行恢复的 airplane 图像。

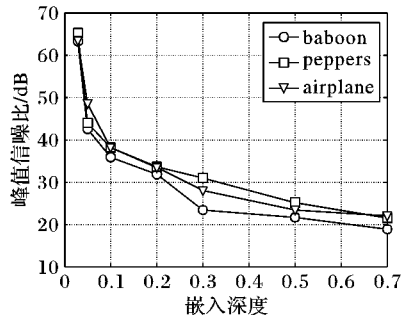
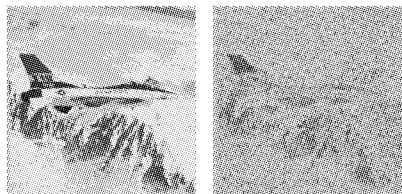


图4 不同控制因子(嵌入深度)下的发布图像峰值信噪比



(a) 正常恢复的图像 (b) 强行恢复的图像
图5 正确恢复和强行恢复的图像

如果图像遭到攻击(比如噪声、裁剪等攻击后),这样会使得最终提取出来的压缩向量发生变化。虽然按照提取算法同样也可以得到一个图像与水印,但是并不能判断得到的是否正确,因此用户恢复时,还需要与原始水印进行比较,如果和原来的水印不一致,则说明得到的不是正确图像。

4 结语

本文在免疫水印算法模型的框架下,提出了一种在小波域上采用可逆隐藏实现可完整精确恢复原始图像的具体算法。通过结合小波变换和可逆水印算法本身所拥有的特点来计算控制因子,达到控制嵌入深度的目的,使得嵌入端可以控制最后图像的失真程度;并且通过置乱加密等安全措施,保证了只有合法的接收端可以获得正确的原始图像。根据该算法

的特点,它可以应用于需要精确图像的领域,如医学图像、军事图像等。

参考文献:

- [1] 彭德云,王嘉祺,王素贞,等. 免疫数字水印技术[J]. 计算机工程与应用, 2006, 19(3): 11-13.
- [2] TIAN J. Reversible data embedding using a difference expansion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(8): 890-896.
- [3] NI Z, SHI Y Q, ANSARI N, et al. Reversible data hiding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(3): 354-362.
- [4] 王俊祥,杨波. 基于直方图平移可逆水印的性能估计[J]. 计算机应用, 2010, 12(12): 3246-3251.
- [5] 王俊祥,倪江群,潘金伟. 一种基于直方图平移的高性能可逆水印算法[J]. 自动化学报, 2012, 38(1): 88-96.
- [6] TAI W L, YEH C M, CHANG C C. Reversible data hiding based on histogram modification of pixel differences[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(6): 906-910.
- [7] TSAI P Y, HU Y C, YEH H L. Reversible image hiding scheme using predictive coding and histogram shifting[J]. Signal Processing, 2009, 89(6): 1129-1143.
- [8] LIN S L, HUANG C F, LIOU M H, et al. Improving histogram-based reversible information hiding by an optimal weight-based prediction scheme[J]. Journal of Information Hiding and Multimedia Signal Processing, 2013, 1(1): 19-33.
- [9] JUNG S W, HA L T, KO S J. A new histogram modification based reversible data hiding algorithm considering the human visual system[J]. IEEE Signal Processing Letters, 2011, 18(2): 721-724.
- [10] WENG S W, PAN J S, GAO X. Reversible watermark combining pre-processing operation and histogram shifting[J]. Journal of Information Hiding and Multimedia Signal Processing, 2012, 3(10): 320-326.
- [11] 李建伟,胡永健,陈开英. 边缘和纹理优先的可逆数据隐藏算法[J]. 计算机应用, 2008, 28(S1): 76-79.
- [12] 周璐,胡永健,曾华飞. 用于矢量数字地图的可逆数据隐藏算法[J]. 计算机应用, 2009, 29(4): 990-993.
- [13] 徐德智,童学锋,宣国荣,等. 基于直方图调整的二值图像无损数据隐藏[J]. 计算机应用, 2009, 29(6): 1651-1653.
- [14] 宣国荣,姚秋明,柴佩琪,等. 基于整数小波阈值嵌入的无损数据隐藏[J]. 计算机应用, 2006, 26(12): 2891-2893.
- [15] CHAN Y K, CHEN W T, YU S S, et al. A HDWT-based reversible data hiding method[J]. The Journal of System and Software, 2009, 82(3): 411-421.

(上接第2197页)

- [4] 李泽安,陈建平,赵为华. 高维数据挖掘中基于中位数回归的特征提取新方法[J]. 计算机应用研究, 2013, 30(2): 374-376.
- [5] LEE M. Mode regression[J]. Journal of Econometrics, 1989, 42(3): 337-349.
- [6] SCOTT D. Multivariate density estimation: theory, practice and visualization[M]. New York: Wiley, 1992.
- [7] YAO W, LINDSAY B, LI R. Local modal regression[J]. Journal of Nonparametric Statistics, 2012, 24(3): 647-663.
- [8] TIBSHIRANI R. Regression shrinkage and selection via the LASSO[J]. Journal of the Royal Statistical Society: Series B, 1996, 58(1): 267-288.
- [9] EFRON B, HASTIE T, JOHNSTONE I, et al. Least angle regression[J]. The Annals of Statistics, 2004, 32(2): 407-489.
- [10] ZOU H. The adaptive LASSO and its oracle properties[J]. Journal

- of the American Statistical Association, 2006, 101(476): 1418-1429.
- [11] FAN J, LI R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456): 1348-1360.
- [12] LI J, RAY S, LINDSAY B. A nonparametric statistical approach to clustering via mode identification[J]. Journal of Machine Learning Research, 2007, 8(8): 1687-1723.
- [13] 潘锋,王建东,牛奔. 基于谱分析的无监督特征选择算法[J]. 计算机应用, 2011, 31(8): 2109-2114.
- [14] 李泽安. 高维数据挖掘中基于正则化估计的特征提取算法[J]. 合肥工业大学学报: 自然科学版, 2012, 35(12): 1655-1658.
- [15] 李泽安,葛建芳,章雅娟. Beta 回归模型在数据挖掘预测中的应用[J]. 南通大学学报: 自然科学版, 2009, 8(3): 83-85.