文章编号: 1000-5862(2012)02-0131-04

高维数据分类中的特征降维研究

刘立月[1], 黄兆华[1], 刘遵雄[2]

(1. 华东交通大学软件学院, 江西 南昌 330013　2. 华东交通大学信息工程学院, 江西 南昌 330013)

摘要: 以高维分类为目标, 从分类的准确率与模型解释性角度探讨了降维的必要性, 分析了特征选择与抽取 2 类方法特点, 并对常用的特征抽取方法, 包括主成分分析(PCA)、偏最小二乘(PLS)和非负矩阵分解(NMF)进行了阐述. 考虑到约减后的数据缺乏稀疏性与可解释性, 提出了基于稀疏正则化的特征抽取模型, 为高维特征降维提供了一种新思路.

关键词: 高维数据; 降维; 特征抽取; 稀疏正则化
中图分类号: TP 181　　文献标志码: A

0　引言

在过去的10年中, 科学技术的进步带来了数据维数爆炸性增长, 成千上万的变量(特征)数目远远超出观测值的数量. 此类问题已广泛地影响到各个领域, 如图像处理、基因微矩阵研究、文本数据分析等. 伴随着数据维数的增长, 模型选择、参数估计、目标函数优化将变得越来越棘手, 已成为统计分析、机器学习、图像处理、模式识别等领域目前面临的普遍现象, 即维数灾难. 维数灾难带来的问题主要表现在3个方面: ①数据维数的增加导致空间数据点更加孤立, 参数空间的全局优化越来越困难; ②高维数据含有更高的噪音, 干扰变量或噪音变量(noisy variables)可能使得原始数据结构更复杂, 隐蔽性更强[1]. 假若信噪比太小, 由于噪音的积累, 对总体均值矢量评估、线性判别规则并不比随机猜测强[2]. ③计算机的运算及存储能力目前已足够强大, 但高维数据处理所需要的存储空间与运算能力仍不可忽视.

处理高维分类问题的自然想法就是首先将数据维数降到合适的大小, 尽可能保持原有数据分类信息, 然后对约减的数据采用标准的分类模型. 考虑到在中、低维数据空间有许多执行得较好的分类技术, 本文重点研究以分类为目标的高维数据特征降维技术.

1　分类与降维

在分类方面, 人们更多地关注的是分类准确率与模型解释性[3]. 分类准确率指的是模型的预测能力; 而模型解释性指的是模型的简洁性, 即提供简单明晰的输入与输出之间的逻辑关系.

在高维数据分析中, 从分类准确率角度来看, 随着变量数目的增加, 尤其当输入变量(m)多于观测值(n)时, 一些传统的分类方法, 如 Fisher 判别将不再合适. 当应用 Fisher 判别规则开展总体均值与协方差矩阵评估时, 由于 $n<m$, 协方差矩阵变得奇异, 因而不稳定; 另一方面, 尽管总体均值矢量的各个组成部分可能获得准确估计, 但积累的评估误差也是很大, 这些评估误差将严重破坏分类准确率. 其它的方法, 如 NN(neural networks)、KNN(k-nearest neighbors)、SVM (support vector machines), 并没有明确要求 $n>m$, 但当 m 很大时, 其分类准确率也表现不尽人意. 从模型解释性角度来看, 输入变量与响应变量应有简单关联性, 因为简单的模型通常能建立一个明晰的数据逻辑结构, 从而使数据有更好的理解性, 如果数据维数太高, 显然很难实现. 降维是高维分类面临的首要任务.

2　降维技术

高维特征空间中, 特征之间可能是冗余的或者

收稿日期: 2012-01-18
基金项目: 国家自然科学基金(61065003,61165004)和江西省教育厅科学技术研究(GJJ12308)资助项目.
作者简介: 刘立月(1970-), 男, 安徽安庆人, 副教授, 硕士, 主要从事机器学习、嵌入式开发方面的研究.

不相关的, 造成高维空间处理的不便, 容易出现过学习现象, 时间与空间开销大, 在不影响分类精度情况下, 需要进行特征降维[4]. 特征降维一般分为2类: 特征选择和特征抽取[5].

特征选择就是从原始特征集中选择一个包含若干显著特征项的真子集, 选择后的特征集大小远比原始特征集小. 特征选择直接选出原始特征集的子集, 不会改变原始空间的性质, 并且特征代表的意义也未改变, 具有直观的解释作用. 常用的特征选择方法有最优子集法、逐步前向或后向选择法、前向逐段回归(forward-stagewise regression)[6]等. 特征抽取是从原始特征空间到新特征空间的一种映射或变换, 抽取后得到的特征含有特征项之间的语义信息、相关性信息, 是对原始特征集在特征相关性层面的压缩. 特征抽取基于原始特征集, 通过变换生成新的组合特征集, 而组合特征往往仅具有数学意义[7]. 特征选择和特征抽取都可以降低特征空间的维数, 从而达到降低计算复杂度和提高分类的准确率的目的, 并为后续分类器的设计提供参数.

在高维的特征空间, 特征选择有时因巨量的计算或其他原因而出现困难, 如对于n个备选特征, 如果采用穷举法需要穷举2^n种情况来搜索使得分类性能最好的一种; 特征选择的另一个不足之处是它的不稳定性[8], 即变量选择的结果会由于数据集合的微小变化而发生大的变化. 当前研究较多的是特征抽取, 常用的特征抽取方法有主成分分析(principal component analysis, PCA)、偏最小二乘(partial least squares, PLS)和非负矩阵分解(nonnegative matrix factorization, NMF)等.

3 特征抽取方法

3.1 主成分分析(PCA)

PCA是一种掌握事物主要矛盾的统计分析方法, 它可以从多元事物中解析出主要影响因素, 揭示事物的本质, 简化复杂的问题. 假设给定n个变量的m个观察值, 形成一个$n\times m$的数据矩阵. PCA 的目标是寻找 r ($r<n$) 个新变量, 使它们反映事物的主要特征, 压缩原有数据矩阵的规模. PCA 分析步骤可概括如下[9-10]: ①将原始数据矩阵进行中心化与标准化预处理; ②构造新的协方差矩阵; ③计算协方差矩阵的特征值与特征向量, 并将特征值按从大到小排列; ④计算主成分贡献率及累计贡献率, 根据累计贡献率要求(一般70% 以上), 选择前面的r个特征向量就能近似表示原始的数据; ⑤计算主成分载荷.

PCA是一种通过降维技术把多个变量化为少数几个主成分(即综合变量)的统计分析方法. 这些主成分能够反映原始变量的绝大部分信息, 它们通常表示为原始变量的某种线性组合(要求方差最大), 同时它们之间又是彼此独立的(主成分间独立). 通过主成分分析, 压缩数据空间, 将多元数据的特征在低维空间里直观地表示出来, 既有利于达到降维, 又有利于主成分的解释.

3.2 偏最小二乘(PLS)

PLS是一种新型的对多组变量进行建模的多元统计数据分析方法[11], 它既是一种特征抽取方法, 又是一种回归分析模型. PLS方法研究的焦点是通过抽取潜在成分, 对包含多自变量和多因变量的数据进行建模分析. 其核心假设是: 认为观测到的数据是由少量潜在成分(不是直接观察或测量到的变量)驱动的系统或进程产生的. 因此, PLS方法研究的关键点是如何得到这些潜在成分.

考虑p个因变量 $y_1,y_2,\cdots,y_p$ 与m个自变量 $x_1,x_2,\cdots,x_m$ 的回归建模问题. PLS回归分析的基本方法为: 首先在自变量集中提取第一成分t_1(t_1是 $x_1,x_2,\cdots,x_m$ 的线性组合, 且尽可能多地提取原自变量集中的变异信息); 同时在因变量集中也提取第一成分u_1, 并要求t_1与u_1相关程度达到最大. 然后建立因变量 $y_1,y_2,\cdots,y_p$ 与t_1的回归, 如果回归方程已达到满意的精度, 则算法终止. 否则继续第二对成分的提取, 直到能达到满意的精度为止. 若最终对自变量集提取r个成分 $t_1,t_2,\cdots,t_r$, PLS将建立 $y_1,y_2,\cdots,y_p$ 与 $t_1,t_2,\cdots,t_r$ 的回归式, 然后再表示为 $y_1,y_2,\cdots,y_p$ 与原自变量的回归方程式, 即偏最小二乘回归方程式[12].

PLS是建立在PCA基础上的一种多对多线性回归建模的方法, 特别当2组变量的个数很多, 且都存在多重相关性, 而观测数据的样本量又较少时, 用PLS回归建立的模型具有传统的经典回归分析等方法所没有的优点.

3.3 非负矩阵分解(NMF)

NMF是一种近来十分流行的非负多元数据描述方法, 常用于维数约减、特征抽取. 假设非负特征矩阵$\boldsymbol{V}$是一个$m\times n$矩阵, m 表示样本数, n 表示特征维数. $\boldsymbol{V}$可以近似地分解为2个非负矩阵$\boldsymbol{W}_{m\times r}$与$\boldsymbol{H}_{r\times n}$的

乘积, 使得

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}, \tag{1}$$

其中r为压缩后的维数, 通常比m和n小得多, 从而$\boldsymbol{W}$和$\boldsymbol{H}$比原矩阵$\boldsymbol{V}$小得多.

NMF算法的基本思想[13]为: 合理地构造一个目标函数, 以此交替地优化$\boldsymbol{W}$和$\boldsymbol{H}$, 从而得到NMF的一个局部最优解. 求解NMF模型的乘性迭代规则使用最为普及.

NMF分解后的所有分量均为非负值(要求纯加性的描述). 从直观理解的角度看, NMF反映了整体是由部分组成的; 从多元统计的角度看, NMF是在非负等限制下, 在尽可能保持信息不变的情况下, 将高维的随机模式简化为低维的随机模式$\boldsymbol{H}$, 而这种简化的基础是估计出数据中的本质结构$\boldsymbol{W}$; 从维数约减的角度看, 因为基矩阵$\boldsymbol{W}$和系数矩阵$\boldsymbol{H}$同时由NMF来确定, 系数矩阵$\boldsymbol{H}$并非为数据矩阵$\boldsymbol{V}$在$\boldsymbol{W}$上的投影, 所以NMF 实现的是非线性的维数约减[14].

4　稀疏特征抽取技术

4.1　特征抽取分析

NMF的分解结果具有一定程度的稀疏性. 使用NMF对数据样本矩阵进行分解产生的特征矩阵$\boldsymbol{W}$具有明显解释意义. 样本类别可以确定为最大编码系数对应的特征向量代表的类别, 实现数据降维和分类一次性完成. NMF能无监督地导致相对稀疏的或局部化的描述, 这是它最主要的特点之一, 但有时它导致的描述的稀疏程度并不能令人满意, 因此稀疏性增强的NMF算法被广泛研究.

PCA可通过对特征矩阵进行奇异值分解 (SVD)求得投影向量. PLS是集PCA、典型相关分析(CCA)和线性回归分析方法的基本功能于一体的多元统计数据分析方法, 能避免自变量之间存在严重多重相关性带来的过拟合问题, 克服了PCA不能辨识噪音与信息的缺点[15]. 对高维数据矩阵使用PCA、PLS进行特征抽取求得的投影向量正交而稠密, 无法给出准确的潜在语义解释, 同时数据分解产生的负值缺乏物理意义. 引入统计学习理论中的正则化技术, 可以使得模型拟合参数值变小, 有助于抑制过学习现象, 可提高分类器的准确率与鲁棒性, 并能实现特征提取任务. 另一方面, 特征项的约减也有利于模型的解释性.

4.2　正则化约束

所谓正则化方法就是在求解模型的优化函数中加上关于参数的惩罚项, 使模型满足一定的性能要求, 其一般形式可表达为

$$L(\beta) + p_\lambda(|\beta|), \tag{2}$$

其中第1 部分为损失函数(loss function), 表示模型拟合的优良性, 如线性回归模型($\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$)常用的平方误差损失函数 $L(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^2$, 第 2 部分为正则化约束(regularization term), 正则化约束意味着对所有特征因子惩罚, 这使得某些β变成 0, 实现相关或重要特征的压缩. 常用的约束函数有: (1) L_0 惩罚, 对应于 AIC 及 BIC 准则 (即以模型中非零系数个数作为惩罚). (2) 桥回归(bridge regression)惩罚: $p_\lambda(|\beta|) = (\lambda|\beta|)^q (0 < q \leqslant 2)$: ①当 q=1 时, 即为 L_1 惩罚, 产生软门限规则, 等同于 Lasso(least absolute shrinkage and selection operator); ②当 q=2 时, 即为 L_2 惩罚, 产生岭回归(ridge regression)规则.

不同的惩罚函数, 得到的解及惩罚效果也不一样. 基于 AIC 及 BIC 准则的最优子集选择相当于 L_0 惩罚; L_2 惩罚能够抑制模型的过拟合问题, 其产生的岭回归存在解析解, 是连续型方法, 能有效克服变量间的高度相关性, 提高预测精度. 岭回归与常规最小二乘法一样无法缩小变量集, 模型的解释性不好; L_1 惩罚能产生最小绝对缩减和变量选择算子(Lasso)回归, 在进行模型参数估计的同时可实现变量选择, 使得部分对模型影响较小的指标变量的系数变为 0, 从而实现模型的变量选择, 缩小了模型的变量集. 由于 L_1 惩罚适合各种模型变量选择, 因此 Lasso 及其相关方法在多元分析领域得到了广泛的运用[16-17].

4.3　正则化的特征抽取

基于正则化约束模型, 可推广应用常规特征抽取方法. 在(2)式中, 以$L(\beta)$为主成分的求解函数, 采用合适的正则化技术, 如L_1惩罚, 将会得到相应的稀疏主成分(sparse PCA)模型. SPCA的求解可以有效地转化为惩罚回归问题. 而Lasso回归问题一般又可以通过最小角回归(least angle regression, LARS)或坐标下降(coordinate descent, CD)算法[18-19]来解决. 因此, SPCA计算也可以利用LARS或CD算法方便地给出.

不同于PCA, PLS考虑了数据类别属性进行特征

抽取，而且能很好地处理多重相关性(即特征空间存在相关性)问题. 借鉴SPCA的实现方法，可以建立稀疏偏最小二乘(sparse PLS)模型. 在NMF优化准则函数中，加入相应的惩罚项，借助迭代非负最小二乘算法，可建立稀疏非负矩阵分解(Sparse NMF)模型.

5 结束语

高维数据在分类之前通常需要使用特征抽取技术进行预处理. 传统的特征抽取方法，包括 PCA、PLS 和 NMF，在高维数据的约减过程中缺乏数据的稀疏性与可解释性，引入正则化约束，建立基于稀疏惩罚的特征抽取模型可为该类问题提供一种有效的解决方案. 基于稀疏正则化的特征抽取方法，不仅适合高维分类问题降维，同时也可应用到高维回归领域.

6 参考文献

[1] Donoho D L. High-dimensional data analysis:the curses and blessings of dimensionality [EB/OL]. [2011-10-16].http://www-stat.stanford.edu/~donoho/Lectures/CBMS/Curses.pdf,2000.

[2] Fan Jianqing, Fan Yingying. High dimensional classification using features annealed independence rules [J]. Annals of Statistics, 2008, 36(6): 2605-2637.

[3] Siva Tian T. Dimensionality reduction for classification with high-dimensional data [D]. California: University of Southern California, 2009.

[4] 奉国和，郑伟. 文本分类特征降维研究综述 [J]. 图书情报工作, 2011(9):.

[5] 陈涛，谢阳群. 文本分类中的特征降维方法综述 [J]. 情报学报, 2005, 24(6): 690-695.

[6] Hastie T, Tibshirani R, Friedman J H. The elements of statistical learning: data mining,inference, and prediction [M]. 2nd ed . New York: Springer, 2009.

[7] 胡洁. 高维数据特征降维研究综述 [J]. 计算机应用研究, 2008, 25(9): 2601-2606.

[8] Breiman L. Heuristics of instability and stabilization in model selection [J]. Annals of Statistics, 1996, 24: 2350-2383.

[9] Jolliffe I T. Principal Component Analysis [M]. 2nd ed. New York: Springer, 2002.

[10] Shen H P, Huang Jianhua. Sparse principal component analysis via regularized low rank matrix approximation [J]. Journal of Multivariate Analysis, 2008, 99: 1015-1034.

[11] Abdi H. Partial least squares regression and projection on latent structure regression [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010(2): 97-106.

[12] 高惠璇. 两个多重相关变量组的统计分析(3)(偏最小二乘回归与 PLS 过程) [J]. 数理统计与管理, 2002, 21(2): 58-64.

[13] Lee D D, Seung H S. Algorithms for non-negative matrix factorization [C]. Cambridge: MIT Press, 2000: 556-562.

[14] 李乐，章毓晋. 非负矩阵分解算法综述 [J]. 电子学报, 2008, 36(4): 737-744.

[15] 王惠文，张志慧，Tenenhaus M. 成分数据的多元回归建模方法研究 [J]. 管理科学学报, 2006, 9(4): 27-32.

[16] Fan Jianqing, Li Runze. Variable selection via nonconcave penalized likelihood and its oracle properties [J]. Journal of American Statistical Association, 2001, 96: 1348-1360.

[17] 王大荣，张忠占. 线性回归模型中变量选择方法综述 [J]. 数理统计与管理, 2010, 29(4): 615-627.

[18] Efron B, Hastie T, Johnstone I, et al. Least angle regression [J]. Annals of Statistics, 2004, 32(2): 407-499.

[19] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent [J]. Journal of Statistical Software, 2010, 33(1): 1-22.

The Research on Dimensionality Reduction for High-Dimensional Data Classification

LIU Li-yue[1], HUANG Zhao-hua[1], LIU Zun-xiong[2]

(1. School of Software, East China Jiaotong University, Nanchang Jiangxi 330013, China;
2. School of Information, East China Jiaotong University, Nanchang Jiangxi 330013, China)

Abstract: With the goal of high-dimensional classification, dimension reduction is discussed from the perspective of classification accuracy and model interpretation, and feature selection and feature extraction characterizes are analyzed. This paper introduces common feature extraction methods, including Principal Component Analysis, Partial Least Squares and Nonnegative Matrix Factorization. Considering the reduced data lacking in sparseness and interpretation, a sparse regularization based feature extraction framework has been proposed, and it provides dimensionality reduction in high-dimensional space a novel and available approach.

Key words: high-dimensional data; dimension reduction; feature extraction; sparse regularization

(责任编辑：冉小晓)