Statistics and Application 统计学与应用, 2020, 9(2), 256-264
Published Online April 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.92028

Hans汉斯

PCA and CUR Decomposition Analysis in Dimensionality Reduction of High-Dimensional Data

Liping He

North China University of Technology, Beijing
Email: 2040345359@qq.com

Received: Mar. 26th, 2020; accepted: Apr. 9th, 2020; published: Apr. 17th, 2020

Abstract

In the era of big data, there are many ways to reduce the dimension of high-dimensional data. Among them, the linear dimension reduction method is the most typical method in PCA and CUR decomposition method, but at present, using the two methods on the research achievements of high-dimensional data dimension reduction is not enough. Therefore, through the discussion and research of the principal component analysis and CUR decomposition method, this paper analyzes the use conditions and practical effects of the two methods. It is concluded that: with the traditional principal component analysis method of the matrix decomposition, in terms of feature selection, CUR decomposition method not only has high accuracy, but also has good interpretability. In terms of matrix recovery, CUR matrix decomposition method has high stability and accuracy, and its accuracy can sometimes reach more than 90%. Therefore, I think CUR matrix decomposition has good application value, and it is worth using CUR matrix decomposition method to reduce the dimension of high-dimensional data.

Keywords

Principal Component Analysis, Line and Column Joint Algorithm, Feature Selection, Matrix Recover

高维数据降维中的PCA与CUR分解对比分析

何丽萍

北方工业大学，北京
Email: 2040345359@qq.com

收稿日期：2020年3月26日；录用日期：2020年4月9日；发布日期：2020年4月17日

摘　要

在大数据的时代，使高维数据降低维度有很多种方法。其中，在线性降维方法中最典型的方法是PCA和

文章引用: 何丽萍. 高维数据降维中的 PCA 与 CUR 分解对比分析[J]. 统计学与应用, 2020, 9(2): 256-264.
DOI: 10.12677/sa.2020.92028

CUR分解方法，但是目前我们利用这两种方法对高维数据降维的研究成果还不够，因此，本文通过对这主成分分析和CUR分解方法的探讨和研究，分析了这两种降维方法的使用条件和实际效果，得出：与传统的主成分分析的矩阵分解方法相比较，在特征选择方面，CUR分解方法不仅具有很高的准确度，而且还具有很好的可解释性；在矩阵恢复方面，CUR矩阵分解方法具有很高的稳定性同时还具有很高的准确度，其准确度有时候能够达到90%以上，因此我认为CUR矩阵分解具有很好的应用价值，值得我们利用CUR矩阵分解法去对高维数据进行降维。

关键词

主成分分析，行列联合算法，特征选择，矩阵恢复

1. 引言

在高维空间中，由于高维数据难以处理，所以表示数据之间相似性度量的 L 距离将会失去意义。在高维数据里存在很多空值，因此数据实际的维度比原始的数据维度小得多，我们可以通过降维的手段转换到低维空间进行处理。有很多种处理高维数据的方法，常见的方法有 PCA 分解、CUR 分解和 SVD 分解等等，在传统的矩阵分解中，虽然能够将高维的数据分解成低维的数据，但是分解出来的结果可解释性不够，主要是因为分解出来的数据不是原始的数据，我们不能用现实生活中的概念去解释结果，只能理解为潜在语义空间，因此分解后矩阵解释性不高。除此之外，由于分解后矩阵的数值不是原来的数据，通过矩阵恢复来预测矩阵稳定性和准确性也很差。本文运用行列选择算法来实现特征选择算法，并且用改进的 CUR 算法来构造恢复矩阵，CUR 分解是在原始数据矩阵中根据概率的大小来选取部分行和列，然后再构造矩阵的分解方法[1]。CUR 分解由原始数据构造而来，其得到的矩阵稀疏且物理意义明确，同时，CUR 分解的算法较为简单，避免了对高维矩阵进行特征值求解，因此其效率也较高，本文主要对这两种降维方法进行对比分析。

矩阵恢复是最近几年非常流行处理高维数据的方法，在一些图像处理方面的应用是非常广泛的，通过分析这几年的大数据竞赛中我们可以发现，在进行用户偏好预测方面，应用最广泛的还是恢复矩阵，本文主要还是利用改进后的恢复矩阵来求解恢复矩阵，其预测的准确性和解释性相对于 PCA 和 SVD 传统的分解方法都很高[2]。

本文有创新点也有不足，创新点在于通过改进的 CUR 分解方法和 PCA 分解方法对高维数据的处理结果可知，CUR 分解方法解释性很高，而且对乳腺癌数据的良性和恶性肿瘤的发病率有很好的预测，可以帮助一些癌症患者去了解发病率以及自己恢复良好的概率，不足的就是数据不是最新的数据，结果相对于现实还是会存在一点误差，我认为其误差可以忽略不计。

2. 理论模型与方法论

2.1. 理论模型

本文所用的数据都是没有缺失值的数据，对有空值的数据去空值，然后对处理后的数据进行标准化处理。在完成上述两个步骤之后，为了防止建立的回归模型出现过度拟合现象，本次研究将数据按照 3:1

的比例分为训练集和测试集，其中训练集用来做回归模型，测试集用来交叉验证选择最终模型以及对模型进行评估，结合 PCA 算法可得。

针对一个给定的训练集，在给定的光滑系数下，Kaiser-Harris 准则建议，选取不同的主成分进行拟合，得出一组模型，记号如式所示[1]

$$\left\{(\lambda)W_I^J\right\}_{i=1,2,3,\cdots,2^n-1;\ j=1,2,3,4,\cdots,m} \tag{1}$$

其中 n 表示在 Kaiser-Harris 准则建议下可以保留的最大主成分个数，

$$\left\{(\lambda)W_I^J\right\}_{i=1,2,3,\cdots,2^n-1;\ j=1,2,3,4,\cdots,m} \tag{2}$$

表示对数据集的第 j 次划分(本次研究仅进行 10 次划分)下，第 i 种情况的拟合模型，此时光滑系数取分别计算出各个模型在验证集中预测的误判率作为经验损失，在经验损失的基础上对模型复杂度(即，采用的主成分个数)施加光滑系数为 λ 的惩罚，得出模型的结构损失。此时对 i 求平均值，得出每种情况预测的平均结构损失，在对 j 取最小，选出平均结构损失最小的情况(即，得出在给定的，平均结构损失最小的模型)，最后在测试集中对进行交叉验证选择出最优的拟合模型作为最终模型[1] [2]。

2.2. 方法论

2.2.1. PCA 方法论

PCA 方法是一类经典的降维方法，其目的在于用一组较少的不相关变量来代替原始数据集中的变量(往往维数较高)，同时保证原始数据集中的信息尽可能多的被解释，称这些提取出来的较少的不相关变量为主成分。其方法论为：最大方差理论。具体表现为寻找一组基，使得所有数据变换为在这组基上的坐标表示后，其方差值最大。形象的可以理解为数据在这组基上的投影尽可能的分能的分散，因此，方法论可以归纳为求解最优化问题：

$$\text{约束优化问题}\begin{cases}\max l^{\mathrm{T}}\Sigma l\\ \text{s.t.}\ l^{\mathrm{T}}l=1\end{cases}, \tag{3}$$

其中 l 是特征根 λ 所对应的单位特征向量，Σ 为协方差矩阵，则最优化问题的解 l 是由协方差矩阵的前 κ 个最大的特征值对应的特征向量构成的[3]。从而 PCA 的输出结果可表示为：

$$Y = xl \tag{4}$$

即，数据维数降低到 κ 维。其中 X 表示降维前的数据，Y 表示降维后的数据。需要注意的是，在 PCA 分析的降维过程中，虽然尽可能的保留了原始信息，但是因为反复对矩阵进行变换，导致提取出的主成分往往难以解释其物理意义，也进一步导致得出的回归模型的解释性不强[3] [4]。

2.2.2. CUR 分析方法

CUR 分解方法主要就是通过将一个矩阵变为三个矩阵 CUR 矩阵，CUR 特征选择方法是计算统计影响力，先计算一下每一列和每一行特征的得分值，然后根据我们已知的得分的大小来决定是否选择这行(列)，通过这一方法，我们可以将不需要的数据去掉，保留有用的数据，得到可以代表其特征的集合，通过这样的集合就可以快速且容易地的分析原数据所具备的特征，本文通过 CUR 矩阵得到恢复矩阵 $\hat{A}$，恢复矩阵： $A \approx CUR = \hat{A}$。

相似程度式[5]：

$$\text{ERROR} = \frac{\|A - CUR\|_F}{\|A\|_F} * 100\% \tag{5}$$

3. 基本算法

3.1. PCA 基本算法

主成分分析算法见表 1 所示：

Table 1. PCA algorithm
表 1. PCA 算法

输入训练数据集 T
输出 PCA 回归模型 算法详细步骤： 1. 记原始数据矩阵为 X，计算数据的协方差矩阵 Σ； 2. 计算协方差矩阵的特征值； 3. 提取较高的几个特征值所对应的特征向量，单位化后构成矩阵记为 μ； 4 对数据进行降维处理，计算 $\mu^{\mathrm{T}}X$； 5. 交叉验证选择最终模型。

3.2. CUR 基本算法

对于给定特征矩阵 $A_{m\times n}$，我们需要首先构建 *CUR* 分解中的低秩矩阵 C，R。在选取过程中我们需要依概率选取更重要的行和列来构造我们的 C，R，因此需要度量所有行和列重要性的指标，我们叫它原始矩阵A的每一列或行的影响力分数，第 i 列的影响力分数记为 $q_j\left(j=1,2,\cdots,n\right)$，第$i$行的影响力分数记为 $p_i\left(i=1,2,\cdots,m\right)$，下面我们给出原始矩阵 $A_{m\times n}$ 每一行和列的影响力分数的形式化度量：

$$p_i=\frac{\sum_{j=1}^{n}A_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}},\ i=1,2,\cdots,m \tag{6}$$

$$q_j=\frac{\sum_{i=1}^{m}A_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}},\ j=1,2,\cdots,n \tag{7}$$

其中 A_{ij} 表示矩阵第 i 行第 j 列的元素值[5]。

有了上述矩阵行和列的影响力分数的形式化描述，可以给出 *CUR* 算法第一步：行列选择算法，见表 2~4 所示：

Table 2. Row and row selection algorithm
表 2. 行列选择算法

输入：原始特征矩阵 $A_{m\times n}$；
输出：矩阵具有代表性的特征列组成的低秩矩阵 C，以及代表用户特征的实例组成的低秩矩阵 R。 算法详细步骤： 1. 根据公式(1)分别计算 A 每一行被选取的可能性 $p_i, i=1,2,\cdots,m$； 2. 根据公式(2)分别计算 A 每一行被选取的可能性 $q_j, j=1,2,\cdots,n$； 3. 从评分矩阵的所有列中从高到低选择 c 列； 4. 从评分矩阵的所有列中从高到低选择 r 行； 5. 返回 $C\in\mathbb{R}^{m\times c}, R\in\mathbb{R}^{m\times c}$。

Table 3. Construction of algorithm 2 matrix U
表 3. 算法 2 矩阵 U 的构造

输入：训练集，行列的矩阵 R，C
输出： 算法详细步骤： 1. 对 C 作广义逆获得矩阵 C 的广义逆矩阵。 2. 对 R 作广义逆获得矩阵 R 的广义逆矩阵。 3. 通过 C 和 R 矩阵的广义逆矩阵来构造 U 矩阵。 4. 返回：$U \in R_{cxr}$ 。

Table 4. Algorithm 3 USES CUR for feature selection and recovery
表 4. 算法 3 利用 CUR 进行特征选择和恢复

输入：wdbc 数据集，一个误差元素，矩阵的秩 k。
输出：用户特征 C，特征 R，恢复矩阵 $\hat{A}$，$\hat{A} \in R_{m \times n}$ 。 1. 先对 wdbc 数据集进行预处理，将 wdbc 数据集转成评分矩阵 $A \in R_{m \times n}$ 。 2. 对矩阵 A 运行列选择算法 1 构造矩阵 C。 3. 对矩阵 A 运行列选择算法 1 构造矩阵 R。 4. 利用算法 2 计算矩阵 U。 5. 计算矩阵 $\hat{A}$，$\hat{A}=CUR$，其中 $\hat{A}$ 被称为对 A 的恢复，$\hat{A}$ 与 A 的偏差的多少决定矩阵恢复效果的好坏。 返回：$C \in R_{m \times c}$，$U \in R_{cxr}$，$R \in R_{r \times n}$，得到恢复矩阵 $\hat{A}$，$\hat{A}=CUR$ 。

4. 实证分析

4.1. 研究对象选取

本文通过 UCI 网站收集到乳腺癌数据集(简称 wdbc 数据集)，其中有 569 的观测值，30 个预测变量，2 个分类，B = 恶性，M = 良性，通过对数据进行处理，分为训练集，测试集，通过 PCA 和 CUR 分解进行降维处理。

4.2. PCA 算法实现

本文选取原始数据的一部分作为训练集和测试集，对训练集做主成分回归得到：

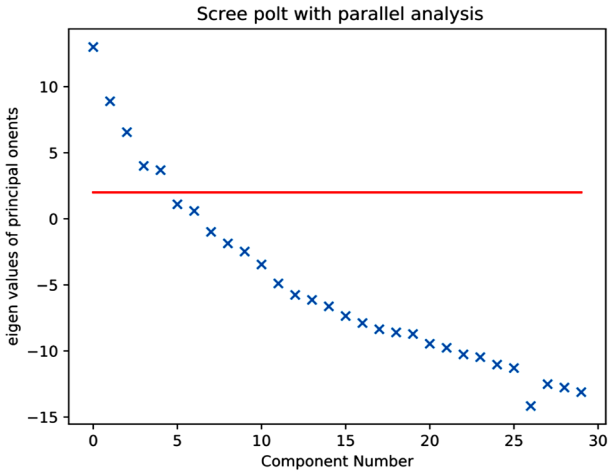

Figure 1. Principal component diagram
图 1. 主成分图

通过观察上图 1 可以发现有 5 个主分依据 Kaiser-Harris 准则建议保留特征值大于 1 的主成分，故而结合图 1 可知，本次研究选择 5 个主成分即可保留原始数据集的大部分信息，但是最终模型是否 5 个主成分均需要保留在回归模型中，则通过交叉验证选择结构损失函数最小的模型[6]。

结构损失函数表示为在采取 0-1 损失函数的基础上，对模型采用主成分的个数进行惩罚，如式(5)所示：

$$\min_{f\in F}\frac{1}{n}W\left(y_i,f\left(x_i\right)\right)+\lambda\left\|f\right\| \tag{8}$$

其中 n 表示数据观测数，f 表示拟合的模型，F 表示所有可能模型构成的集合，$W\left(y_i,f\left(x_i\right)\right)$ 表示损失函数，$\left\|f\right\|$ 表示模型 f 中采用的主成分个数，λ 表示光滑系数，最终经过交叉验证选定的主成分个数为 4，此时平均误判率为 0.07511737。在将模型代入测试集进行预测后，得到在测试集的误判率为 0.063388028，预测效果还可以，并且在对实验数据的解释度上 85%，即对实验数据的信息丢失度较小[7]。

实验数据的协方差矩阵特征值最大的两个变量对于数据的分类情况如图 2 所示：

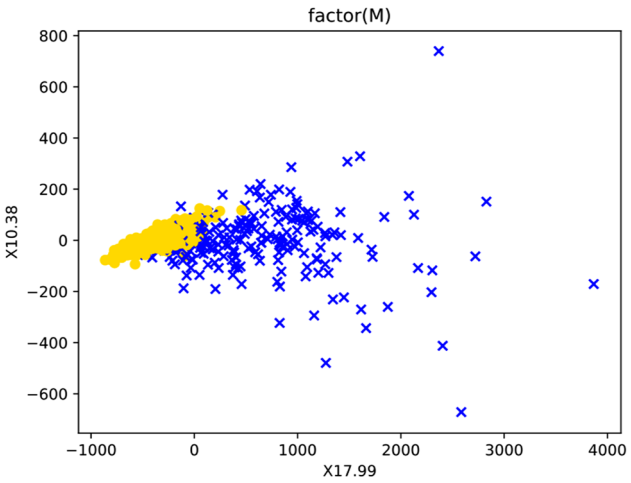

Figure 2. Classification of data by maximum two variables of covariance matrix eigenvalue of experimental data
图 2. 实验数据协方差矩阵特征值最大两个变量对数据的分类情况

同时，第 1 主成分与第 5 主成分对实验数据的分类情况绘图如图 3 所示：

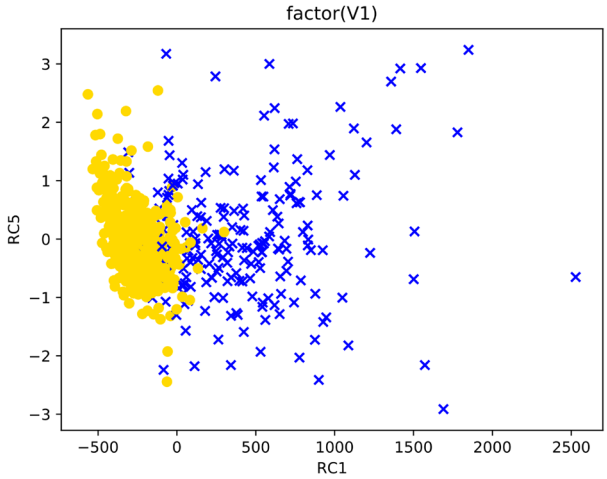

Figure 3. PCA classification of experimental data
图 3. PCA 对实验数据的分类情况

从图 2 与图 3 中可以看出对于本次研究的实验数据主成分分析在一定程度上达到了分类的效果，但是两种类别的分界线明显。

4.3. CUR 矩阵恢复

该实验所用的 wdbc 数据集包含了 569 位患者的乳腺癌疾病诊断数据集，该数据集包含 32 列特征。我们的矩阵相似性评估方法用误差率来度量[8]，表示如下：

$$\mathrm{ERROR}(\%)=\frac{\|A-C\cdot U\cdot R\|_F}{\|A\|_F} \tag{9}$$

其中 $\|A\|_F=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}A_{i,j}^2\right)^{\frac{1}{2}}$。

4.3.1. 实验构造

通过对 wdbc 数据集进行 CUR 矩阵分解可以得到数据集的近似矩阵，我们首先统计 wdbc 数据集的 32 列特征的影响力评分，也就是观察乳腺癌有哪些主要特征，通过 matlab 程序计算出影响力评分见下图 4、图 5 所示，图中横坐标表示特征，纵坐标表示影响力度量值[9]。

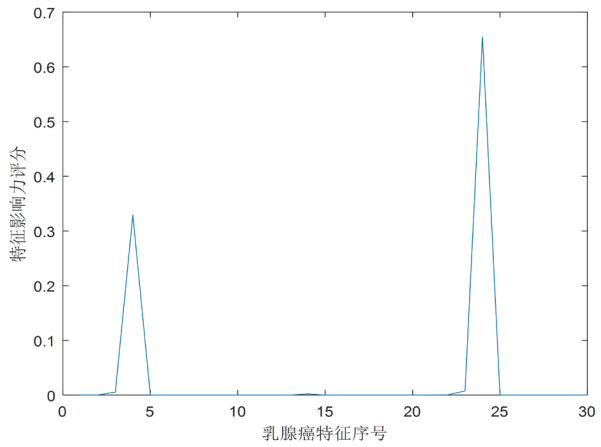

Figure 4. Influence distribution of 32 characteristics of breast cancer
图 4. 乳腺癌疾病 32 列特征的影响力分布

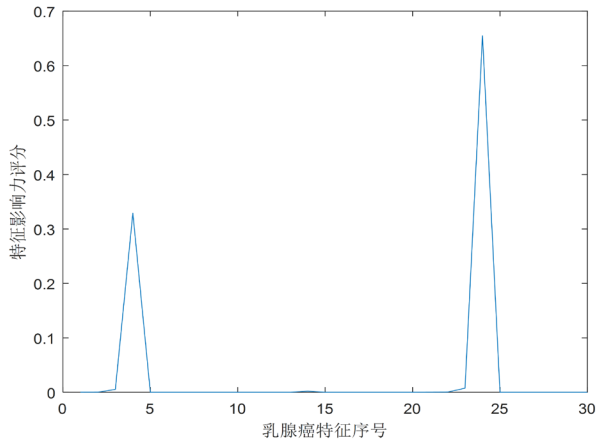

Figure 5. Influence distribution of 569 patients with breast cancer
图 5. 乳腺癌疾病 569 位患者的影响力分布

4.3.2. c 和 r 值的确定

通过 *cur* 算法总能算得给定 c 值和 r 值之后的恢复矩阵 $c\cdot u\cdot r$，再根据误差定量公式总能算得一个恢复矩阵与原始矩阵的误差值[10]，我们首先给定恒定的 r 值，遍历所有的 $c\left(1\le c\le 30\right)$ 值，观察误差值的变化情况如下图所示，其中横坐标是 c 的取值，纵坐标是矩阵误差值：

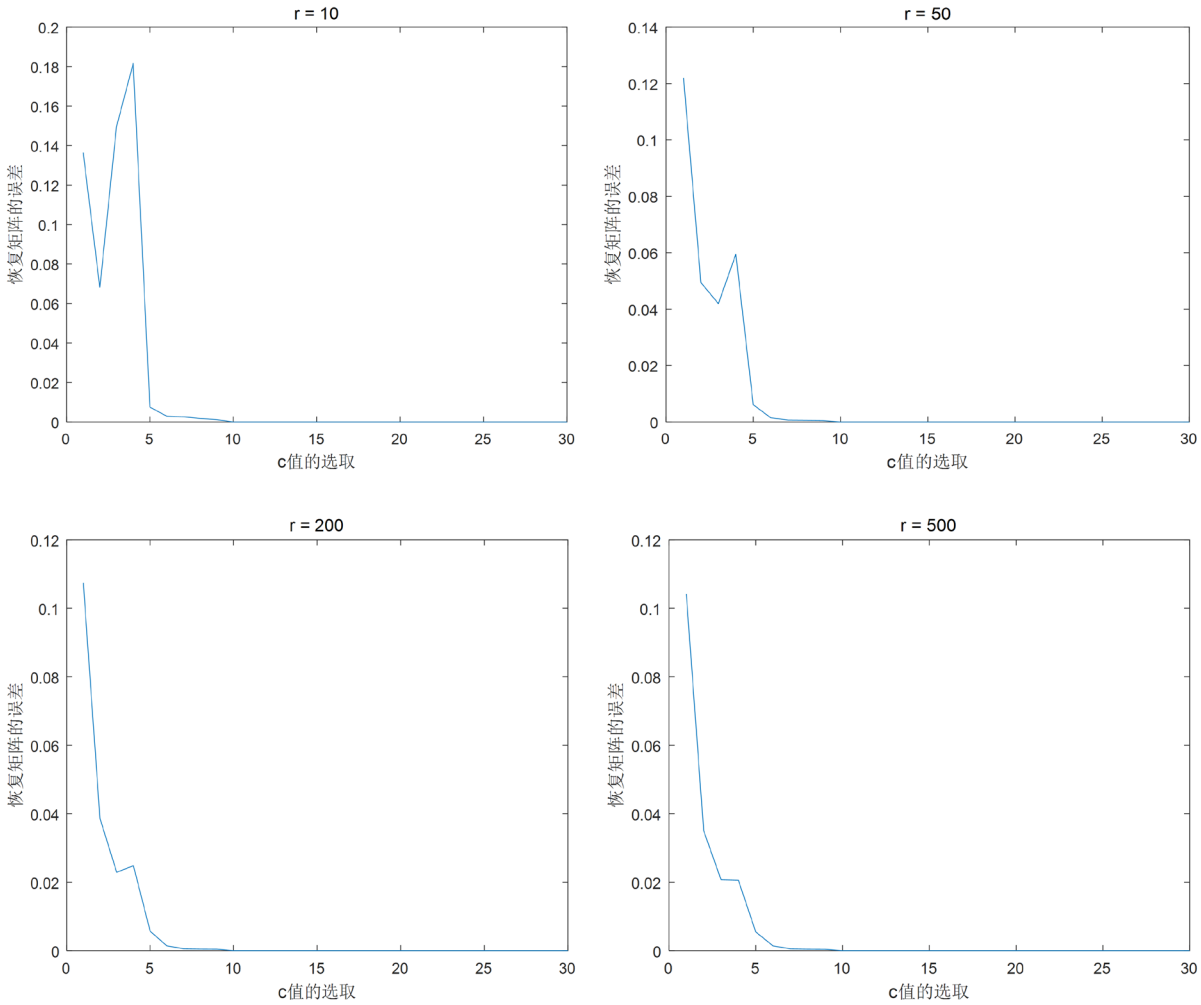

Figure 6. Influence of the selection of constant value on the error of response matrix
图 6. r 值恒定时 c 值的选取对回复矩阵误差的影响

见上图 6 发现 r 值在取不到 15 时，矩阵恢复误差接近 0。通过上述实验探索我们给定 $c=10, r=12$ 算得恢复矩阵误差值为 0.0005014560016610 53%。

5. 结论

通过 PCA 和 CUR 降维方法对乳腺癌数据集进行分类处理，已知 PCA 方法是利用降维的思想，在损失信息很少的前提下，把多个指标转化为几个综合指标的多元统计方法，并且可以消除评价指标之间的相互影响，因为降维，所以也为我们减少了工作量，但是由于主成分不是从原始数据中提取出来的，所以解释性不够好，并且在乳腺癌数据集上进行降维处理发现 CUR 的误差率远小于 PCA，所以在减低误差率这一块 CUR 要好于 PCA。

CUR 算法在选取行列的时候是根据行列的实际情况比较智能地进行选择，在研究大数据时，我们可以理解为在选择行和列的时候，CUR 算法判断矩阵列的统计影响力来判断特征的显著程度，然后根据显著的特征从电影列中进行抓取，所以构造出的 C(R)矩阵也反映了癌症良性恶性的现实特征，并且因为 C(R)矩阵是由真实的行列构成，因此将 C(R)矩阵中的行列数据和 Wdbc 数据集相比对就能够确定细胞核特征和良性恶性的具体特征，但是在软件操作上 CUR 比 PCA 复杂得多，所以如果不是研究高维稀疏的数据集建议用 PCA 降维方法，但在解释性方面，由于 CUR 分解方法是在原始数据集中提取出来的，所以其可解释性很强[11]。

基金项目

北方工业大学毓优人才项目 207051360020XN140/007。

参考文献

[1] Bobadilla, J., Ortega, F. and Hernando, A. (2013) Recommender Systems Survey. *Knowledge-Based Systems*, **46**, 109-132. https://doi.org/10.1016/j.knosys.2013.03.012

[2] 林海明. 对主成分分析法运用中十个问题的解析[J]. 统计与决策, 2007(16): 16-18.

[3] 雷恒鑫, 刘惊雷. 基于行列联合选择矩阵分解的偏好特征提取[J]. 模式识别与人工智能, 2017, 30(3): 279-288.

[4] 雷恒鑫, 刘惊雷. 利用 CUR 矩阵分解提高特征选择与矩阵恢复能力[J]. 计算机应用, 2017, 37(3): 640-646.

[5] Kumar, R., Verma, B.K. and Rastogi, S.S. (2014) Social Popularity Based SVD++ Recommender System. *International Journal of Computer Applications*, **87**, 33-37. https://doi.org/10.5120/15279-4033

[6] 张梦阳. 矩阵分解的常用方法[J]. 成才之路, 2012(36): 39-39.

[7] Jia, Y., Zhang, C., Lu, Q., *et al*. (2014) Users' Brands Preference Based on SVD++ in Recommender Systems. *Proceedings of the 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications*, Ottawa, ON, 29-30 September 2014, 1175-1178.

[8] 王群英. 矩阵分解方法的探究[J]. 长春工业大学学报, 2011, 32(1): 95-101.

[9] Xu, F., Gu, G., Kong, X., *et al*. (2016) Object Tracking Based on Two-Dimensional PCA. *Optical Review*, **23**, 231-243. https://doi.org/10.1007/s10043-015-0178-2

[10] Drineas, P., Mahoney, M.W. and Muthukrishnan, S. (2008) Relative-Error CUR Matrix Decompositions. *SIAM Journal on Matrix Analysis and Applications*, **30**, 844-881. https://doi.org/10.1137/07070471X

[11] 李明. 基于矩阵分解理论学习的数据降维算法研究[D]: [硕士学位论文]. 大连: 辽宁师范大学, 2011.