• 研究简报 •

一种适用于处理中药指纹图谱数据的主成分正交分解算法

朱尔一, 王小如
(厦门大学化学化工学院 现代分析科学教育部重点实验室, 福建 厦门 361005)

摘要: 中药指纹图谱数据具有变量数很大而样本数较小的特点, 本文中采用拉格朗日求极值的方法得到一种新的适合用于处理这类数据的主成分正交分解算法. 结果表明: 所得到新的算法, 在处理中药指纹图谱数据时, 与传统的主成分分析算法比较, 节省存储单元, 计算量小, 计算速度快, 因而计算效率高.

关键词: 主成分分析; 指纹图谱; 判别分析
中图分类号: O 6-04 文献标识码: A 文章编号: 0438-0479(2005)06-0884-02

中药指纹图谱数据的特点是变量的维数很大, 而样本数相对较小, 即每一个样本的指纹图谱数据通常采集 2000 多个数据点, 而对于每一种中药材在很多情况下样本数一般小于 100, 采用一般的主成分正交分解算法处理这类数据计算效率不是很高, 因此有必要发展适合处理这类变量数远大于样本数数据的算法. 在中药指纹图谱数据研究中, 需要根据已有的中药指纹图谱数据, 建立模式识别判别分类模型, 根据计算结果来判别中药药材的种类, 生长方式(野外生长或人工栽培)等. 主成分分析法是分析这类数据并建立这类判别分类模型的一种常见方法, 而主成分正交分解算法是该法的基础.

1 主成分正交分解原理

主成分正交分解算法是一种对数据矩阵 X 实施以下序列的正交变换的过程

$$\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{r}_i \quad (i = 1, 2, \cdots) \tag{1}$$

在变换过程中, 使得到的矢量 $\boldsymbol{t}_i$ 的方差, 即

$$\boldsymbol{t}_i^T\boldsymbol{t}_i = \max \tag{2}$$

为最大值, 这样在经过正交变换得到的相互正交的矢量或主成分 $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots$, 中, 排在前面的矢量 $\boldsymbol{t}_1$ 的方差值为最大, 而排在后面的矢量的协方差值依次减小, 而方差值小的主成分被认为含有较多的噪音的成分, 因此在建模时通过删除方差值小的主成分, 以达到噪音分离的目的.

上述正交变换中还需满足以下约束条件

$$\boldsymbol{r}_i^T\boldsymbol{r}_i = 1 \quad (i = 1, 2, \cdots) \tag{3}$$

即矢量 $\boldsymbol{r}_i$ 为单位矢量.

另外各主成分相互正交或两两正交

$$\boldsymbol{t}_i^T\boldsymbol{t}_j = 0 \quad (i \neq j) \tag{4}$$

因此求解主成分正交分解算法中的变换矢量 $\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots$, 从而获得主成分 $\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots$, 可看作一个求解在约束条件(3)和(4)下的极值问题(2)[1~2].

2 主成分正交分解算法

采用拉格朗日乘子法, 求解上述约束条件下的极值问题(2), 可得到传统主成分正交分解迭代算法, 见表 1.

表 1 主成分正交分解算法

Tab. 1 The orthogonal expansion algorithm of principal component

1. 计算方差矩阵	$(\boldsymbol{X}^T\boldsymbol{X})_1 = \boldsymbol{X}^T\boldsymbol{X}$
2. 迭代计算	$\boldsymbol{r}_i = 1/\lambda_i(\boldsymbol{X}^T\boldsymbol{X})_i\boldsymbol{r}_i$ 选取 λ_i 使矢量 $\boldsymbol{r}_i$ 为单位矢量
3. 计算主成分矢量	$\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{r}_i$
4. 更新方差矩阵	$(\boldsymbol{X}^T\boldsymbol{X})_{i+1} = (\boldsymbol{X}^T\boldsymbol{X})_i - \lambda_i\boldsymbol{r}_i\boldsymbol{r}_i^T$
5. 返回到 2, 重复计算 2~4, 从 $i = 1$ 开始计算, 直到提取所有的有用信息	

在表 1 迭代算法中, 第 2 步的迭代计算, 可采用关系式 $(\lambda^{(k+1)} - \lambda^{(k)})/\lambda^{(k)} < 10^{-8}$ 作为迭代收敛条件.

3 适合处理变量数多样本数少数据的主成分正交分解算法

收稿日期: 2004-11-19
基金项目: 福建省自然科学基金(C0210006)和福建中药 GAP 关键技术研究基金(2002Y024)资助
作者简介: 朱尔一(1957-), 男, 博士, 副教授.

中药指纹图谱数据的特点是变量数很大, 既矩阵 X 的列数很大(有时 > 1 000), 而样本数相对较小, 既矩阵 X 的行数较小(< 100), 对这类变量数多样本数少数据, 以上表 1 迭代算法中, 矩阵$(X^TX)_i$ 的尺寸相对很大, 因此计算量相对很大, 对这类数据, 通过以下改进, 可找到计算效率更高的算法.

用矩阵 X 左乘表一中的第 2 步迭代计算 $r_i = 1/\lambda(X^TX)_i r_i$, 并注意关系式 $t_i = Xr_i$ 可得到公式:

$$\lambda t_i = (XX^T)_i t_i \qquad (5)$$

而表 1 中的第 4 步经过推导可改为以下递推关系式

$$(XX^T)_i = (XX^T)_{i-1} - t_i t_i^T \qquad (6)$$

另外矢量 r_i 可由以下关系式求得

$$\lambda r_i = X^TXr_i = X^Tt_i \qquad (7)$$

以方程(5), (6) 和(7) 为基础的主成分正交分解算法(适合变量多样本少的问题) 见表 2.

表 2　主成分正交分解算法(适合变量多样本少数据)

Tab. 2　The orthogonal expansion algorithm of principal component (for the data set with many variables and few objects)

1. 计算方差矩阵	$(XX^T)_1 = XX^T$
2. 迭代计算	$t_i = 1/\lambda_i(XX^T)_i t_i$ 选取 λ_i 使矢量 t_i 为单位矢量
3. 计算矢量	$r_i = 1/\lambda_i X^T t_i$ 选取 λ_i 使矢量 r_i 为单位矢量 $t_i = Xr_i$
4. 更新方差矩阵	$(XX^T)_{i+1} = (XX^T)_i - t_i t_i^T$
5. 返回到 2, 重复计算 2 ~ 4, 从 $i = 1$ 开始计算, 直到提取所有的有用信息.	

以上表 1 迭代算法与表 2 迭代算法的主要区别是, 采用矩阵$(XX^T)_i$ 来代替矩阵$(X^TX)_i$, 而对于变量数多样本数少的数据, 矩阵$(XX^T)_i$ 的尺寸小于矩阵$(X^TX)_i$ 的尺寸, 因此处理这类数据, 表 2 中的迭代算法的计算效率与表 1 中的迭代算法相比较有较大的提高. 以上两种算法可得到相同的结果.

4　结果与讨论

对一些实际的中药指纹图谱数据的处理表明: 表 1 和表 2 中的算法得到的结果(包括 r_i, t_i, λ_i) 完全相同. 在其中的一套甘草的 HPLC 分析数据处理中, 采用主成分正交分解算法, 对各种类别甘草(拉乌尔甘草, 光果甘草和胀果甘草) 进行分类研究. 在该数据中矩阵 X 中的变量数或列数为 325, 而矩阵 X 中的样本数为 23, 采用表 1 和表 2 中的两种算法计算得到完全相同的 λ 值(极值问题中的极大值), 如下表所示.

表 3　矢量 t_i 的方差 λ_i 值(前 8 个)

Tab. 3　The variance values λ_i of vector t_i (The first 8 values)

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
217 3.2	155 4.3	100 9.2	670.8	438.8	373.4	233.4	189.8

由于矩阵$(XX^T)_i$ 的维数为 23 × 23, 而矩阵$(X^TX)_i$ 的维数为 325 × 325, 表 2 中的算法的速度快得多.

参考文献:

[1] Zhu Eryi, Barnes R M. A simple iteration algorithm for PLS regression[J]. J. of Chemometrics, 1995, 9: 363 – 372.

[2] 朱尔一, 林雍静, 庄峙厦, 等. 一种用于二类样本判别分析的 PLS 方法[J]. 高等学校化学学报, 1997, 18: 212 – 215.

[3] 朱尔一, 扬芃原. 化学计量学技术及应用[M]. 北京: 科学出版社, 2001. 100 – 107.

An Orthogonal Expansion Algorithm of Principal Component Suitable to Deal with the Fingerprinting Data of Chinese Medicine

ZHU Er-yi, WANG Xiao-ru

(College of Chemistry and Chemical Engineering, Key Lab. of Analytical Sciences of the MOE, Xiamen University, Xiamen 361005, China)

Abstract: The fingerprinting data sets of Chinese medicine are the data sets with large number of variables and few objects. A new kind of orthogonal expansion algorithm of Principal Component that is suitable to deal with this kind of data sets has been obtained by use of the Lagrange method of solving extremum problem in this paper. The results indicate that by comparing with the traditional Principal Component Analysis algorithm the new presented algorithm is memory-saving, with small amount of calculation, fast and effective when dealing with the fingerprinting data of Chinese medicine or the data with large number of variables and few objects.

Key words: principal component analysis; fingerprinting; discrimination analysis