

非负矩阵分解及其在模式识别中的应用

刘维湘* 郑南宁* 游屈波†

(西安交通大学人工智能与机器人研究所, 西安 710049. * 同等贡献, † 联系人, E-mail: gbyou@aiar.xjtu.edu.cn)

摘要 矩阵分解是实现大规模数据处理与分析的一种有效工具. 非负矩阵分解(non-negative matrix factorization, NMF)算法是在矩阵中所有元素均为非负的条件下对其实现的非负分解, 这为矩阵分解提供了一种新的思路. 非负矩阵分解方法在智能信息处理和模式识别研究领域具有十分重要的应用意义. 本文介绍非负矩阵分解的基本思想和一些最新的研究成果, 结合研究工作讨论在概率模型的框架下实现非负矩阵分解的目标函数和相应的算法, 以及非负矩阵分解与知觉过程信息处理的关系, 针对模式识别的实际问题给出具体的非负矩阵分解的应用实例, 并提出非负矩阵分解及其应用中有待进一步研究的新问题.

关键词 非负数据 特征提取 非负矩阵分解 入侵检测 数字水印 脑电信号处理

矩阵分解在很多领域获得了广泛的应用^[1], 在数值代数中, 利用矩阵分解可以将规模较大的复杂问题转化为小规模简单子问题来求解; 在应用统计学领域, 通过矩阵分解得到原数据矩阵的低秩逼近, 从而可以发现数据的内在结构特征. 在机器学习和模式识别^[2]的应用中, 矩阵的低秩逼近可以大大降低数据特征的维数, 节省存储和计算资源.

1999 年 Lee 和 Seung^[3]在 *Nature* 上发表了非负矩阵分解算法, 该算法是在矩阵中所有元素均为非负的条件下对其实现非负分解. 从计算的角度来看, 矩阵分解的结果中可以存在负值, 但负值元素在实际问题中往往缺失物理意义. 非负矩阵分解方法则提供了一种新的矩阵分解思路. 由于其分解算法实现简便, 分解的结果中不出现负值, 而且具有可解释性和明确的物理意义, 以及占用存储空间少等优点, 已经引起许多科学家和研究人员的广泛重视. 通常, 信息的表示或编码的各种方法基本上可分为线性和非线性两大类. 线性表示方法是将信号分解为一组基本信号的线性加权组合, 其中基本的信号称为基信号, 对应的权重称为组合系数. 根据基信号的特点, 这种线性表示方法又可以分为两种: 一种是其基信号是固定的, 与处理的数据无关, 如傅里叶分析和小波分析方法; 另外一种基信号与处理的数据相关, 如主成分分析和独立分量分析. 本文介绍的非负矩阵分解方法是后一种表示方法.

在人工智能、机器学习以及计算机视觉和模式识别等研究领域, 为了提高算法或系统的“智能”水平而提出的一些智能计算理论和模型, 大多借鉴于人

类大脑的信息处理机制或对人的认知过程、心理以及生理方面的研究成果. 人类大脑对外界事物整体的感知是建立在对其部分感知基础之上的. 然而, 大脑是如何实现这种对部分的感知? 我们又如何利用计算机来模拟这种功能? 非负矩阵分解方法为解决这类问题提供了一种新的思路. 在非负性约束下, 对非负数据进行非负分解, 采用简单有效的乘性迭代算法, 通过学习得到的基向量中含有关于物体的局部特征的信息. 与传统的主成分分析(PCA)相比, 非负矩阵分解的结果合乎大脑感知的直观体验, 并具有明确的物理含义.

1 非负矩阵分解的定义

信息或信号处理的许多数据具有非负性的特点, 如灰度图像、物质成分含量、文章中单词出现的次数和统计学中的概率转移矩阵等. 在用线性表示方法处理这类数据时, 往往要求分解的结果(包括基向量和系数)都是非负的. 此时若采用传统的因子分析方法, 如主成分分析, 因为其结果中含有负数而失去了物理意义, 而采用非负(正)矩阵分解方法就可以避免这一点^[4].

非负矩阵分解是一种多变量分析方法. 假设处理 m 个 n 维空间的样本数据, 用 $X_{n \times m}$ 表示. 该数据矩阵中各个元素都是非负的, 表示为 $X \geq 0$. 对矩阵 $X_{n \times m}$ 进行线性分解, 有

$$X_{n \times m} \approx B_{n \times r} C_{r \times m}, \quad (1)$$

其中 $B_{n \times r}$ 称为基矩阵, $C_{r \times m}$ 为系数矩阵. 若选择 r 比 n 小, 用系数矩阵代替原数据矩阵, 就可以实现对原

数据矩阵的降维, 得到数据特征的降维矩阵, 从而减少存储空间, 节约计算资源; 从数据编码表示来看, 以基矩阵为码本, 则系数矩阵就为相应的编码系数. 文献[3]从矩阵分解的角度就主成分分析、向量量化、独立分量分析和非负矩阵分解作了讨论.

2 非负矩阵分解的算法

为了实现矩阵的非负分解, 首先需要定义一个损失函数来刻画分解前后的逼近程度, 然后在非负性约束下求解. 最早提出的正矩阵分解方法采用传统的梯度下降算法与加性迭代规则^[4]. 文献[3,5]采用乘性迭代规则, 更适合非负数据的特点, 即, 在非负性初始化的基础上, 在迭代过程中能简单地保持非负性, 而加性迭代规则就需要一个强制将负值变为零的步骤.

2.1 非负矩阵分解的概率模型

将矩阵分解看成如下含加性噪声的线性混合体模型^[6]

$$X_{n \times m} = B_{n \times r} C_{r \times m} + E_{n \times m}, \quad (2)$$

其中 $E_{n \times m}$ 为噪声矩阵. 进一步, 也可以将上式写成

$$X_{ij} = (BC)_{ij} + E_{ij}. \quad (3)$$

为了求解因子矩阵 B, C , 考虑如下的最大似然解^[6]:

$$\begin{aligned} \{B, C\} &= \arg \max_{B, C} p(X | B, C) \\ &= \arg \min_{B, C} [-\log p(X | B, C)], \end{aligned} \quad (4)$$

假设噪声服从不同的概率分布, 就可以得到不同类型的目标函数. 考虑噪声是高斯噪声, 即

$$p(X_{ij} | B, C) = \exp \left\{ -\frac{1}{2} \left[\frac{X_{ij} - (BC)_{ij}}{\sigma_{ij}} \right]^2 \right\} / (\sqrt{2\pi} \sigma_{ij}), \quad (5)$$

其中 σ_{ij} 是对每个观测值给定的权重. 令

$$p(X | B, C) = \prod_{ij} p(X_{ij} | B, C), \quad (6)$$

则最大似然解是最小化如下的损失函数:

$$L(B, C) = \frac{1}{2} \sum_{ij} [X_{ij} - (BC)_{ij}]^2 / \sigma_{ij}^2 + \sum_{ij} \log(\sqrt{2\pi} \sigma_{ij}). \quad (7)$$

若假设 $\sigma_{ij} = 1$ 并忽略因子 $1/2$ 和常数项 $\sum_{ij} \log(\sqrt{2\pi} \sigma_{ij})$,

则得到

$$L_{ED}(B, C) = \sum_{ij} [X_{ij} - (BC)_{ij}]^2. \quad (8)$$

采用传统的梯度法, 有

$$\begin{aligned} \frac{\partial L_{ED}}{\partial B_{ik}} &= -2 \left[(XC^T)_{ik} - (BCC^T)_{ik} \right], \\ \frac{\partial L_{ED}}{\partial C_{kj}} &= -2 \left[(B^T X)_{kj} - (B^T BC)_{kj} \right], \end{aligned} \quad (9)$$

于是可以得到如下的加性迭代规则^[6]:

$$\begin{aligned} B_{ik} &\leftarrow B_{ik} + \phi_{ik} \left[(XC^T)_{ik} - (BCC^T)_{ik} \right], \\ C_{kj} &\leftarrow C_{kj} + \phi_{kj} \left[(B^T X)_{kj} - (B^T BC)_{kj} \right]. \end{aligned} \quad (10)$$

如果设置 $\phi_{ik} = \frac{B_{ik}}{(BCC^T)_{ik}}$, $\phi_{kj} = \frac{C_{kj}}{(B^T BC)_{kj}}$, 式(10)加性迭代规则就变成了如下的乘性迭代规则^[5]:

$$B_{ik} \leftarrow B_{ik} \frac{(XC^T)_{ik}}{(BCC^T)_{ik}}, \quad C_{kj} \leftarrow C_{kj} \frac{(B^T X)_{kj}}{(B^T BC)_{kj}}. \quad (11)$$

如果考虑噪声为泊松噪声, 即有

$$p(X_{ij} | B, C) = \frac{(BC)_{ij}^{X_{ij}}}{X_{ij}!} \exp[-(BC)_{ij}]. \quad (12)$$

同理, 得到其最大似然解是最大化如下的损失函数:

$$L(B, C) = \sum_{ij} [X_{ij} \log(BC)_{ij} - (BC)_{ij} - \log(X_{ij}!)], \quad (13)$$

这实际上等价于最小化如下的广义 KL 距离:

$$L_{KL}(B, C) = \sum_{ij} \left(X_{ij} \log \frac{X_{ij}}{(BC)_{ij}} - X_{ij} + (BC)_{ij} \right). \quad (14)$$

同样, 为了最大化 $L(B, C)$, 采用梯度算法, 有以下的加性迭代规则^[6]:

$$\begin{aligned} B_{ik} &\leftarrow B_{ik} + \phi_{ik} \left[\sum_j C_{kj} \frac{X_{ij}}{(BC)_{ij}} - \sum_j C_{kj} \right], \\ C_{kj} &\leftarrow C_{kj} + \phi_{kj} \left[\sum_i B_{ik} \frac{X_{ij}}{(BC)_{ij}} - \sum_i B_{ik} \right]. \end{aligned} \quad (15)$$

如果设置 $\phi_{ik} = \frac{B_{ik}}{\sum_j C_{kj}}$, $\phi_{kj} = \frac{C_{kj}}{\sum_i B_{ik}}$, 则得到如下的乘

性迭代规则^[5]:

$$\begin{aligned} C_{kj} &\leftarrow C_{kj} \frac{\sum_i B_{ik} X_{ij} / (BC)_{ij}}{\sum_i B_{ik}}, \\ B_{ik} &\leftarrow B_{ik} \frac{\sum_j C_{kj} X_{ij} / (BC)_{ij}}{\sum_j C_{kj}}. \end{aligned} \quad (16)$$

从以上分析可以看到, 当考虑不同的噪声类型时, 可以得到不同的目标函数用来实现矩阵分解. 因此, 我

们用拉普拉斯噪声来得到一类用于实现分解的新目标函数. 当噪声服从拉普拉斯分布时, 可以得到如下似然概率:

$$p(X_{ij} | B, C) = \exp \left\{ - \left| \frac{X_{ij} - (BC)_{ij}}{\sigma_{ij}} \right| \right\} / (2\sigma_{ij}), \quad (17)$$

由此可以得到如下的损失函数:

$$L_l(B, C) = \sum_{ij} \left| \frac{X_{ij} - (BC)_{ij}}{\sigma_{ij}} \right|. \quad (18)$$

我们可以采用线性规划方法求解式(18), 但式(18)的损失函数不是连续可微的, 通常采用如下的近似函数^[7]:

$$L_{app}(B, C) = \sum_{ij} \log \left\{ \cosh \left[\gamma \left(\frac{X_{ij} - (BC)_{ij}}{\sigma_{ij}} \right) \right] \right\}, \quad (19)$$

其中 $\gamma \gg 1$.

综上所述, 我们可以把非负矩阵分解归结为求解如下约束优化问题:

$$\begin{aligned} \min \quad & f(B, C) \\ \text{s.t.} \quad & B \geq 0, C \geq 0. \end{aligned} \quad (20)$$

其中 $f(B, C)$ 为某一损失函数. 在迭代中, 往往还要求对基矩阵的各列向量进行归一化, 如文献[3]中根据 1 范数将基向量进行归一化, 即要求 $\sum_i B_{ik} = 1, \forall k$,

得到如下的乘性迭代规则:

$$\begin{aligned} B_{ik} &\leftarrow B_{ik} \sum_j C_{kj} X_{ij} / (BC)_{ij}, \\ B_{ik} &\leftarrow \frac{B_{ik}}{\sum_l B_{lk}}, \\ C_{kj} &\leftarrow C_{kj} \sum_i B_{ik} X_{ij} / (BC)_{ij}. \end{aligned} \quad (21)$$

另外, 我们探讨了将基向量按 p 范数进行归一化的情形, 并得到了很有意义的结果^[8].

2.2 局部非负矩阵分解(LNMF)

非负矩阵分解只是要求所分解的因子矩阵为非负的, 如果对分解的因子矩阵考虑更多的约束, 就可以强化分解的结果. Li 等人^[9]正是基于这个想法, 从如下 3 个方面给非负矩阵分解的因子施加约束: 要求一个基向量的各个成分不应该被分解得太多, 同时用来表示原始数据的基向量的成分数目尽可能少; 要求不同的基向量尽可能地接近正交化; 要求含有重要信息的成分被保留. 在此基础上, 他们提出的局部非负矩阵分解问题就转化为如下的约束优化求解

问题:

$$\begin{aligned} \min L_{LNMF}(B, C) &= \sum_{ij} \left(X_{ij} \log \frac{X_{ij}}{(BC)_{ij}} - X_{ij} + (BC)_{ij} \right) \\ &\quad + \alpha \sum_{ij} U_{ij} - \beta \sum_i V_{ii} \\ \text{s.t.} \quad &B \geq 0, C \geq 0, \sum_i B_{ik} = 1, \forall k, \end{aligned} \quad (22)$$

其中 $U = B^T B$, $V = CC^T$. 其求解的乘性迭代规则为^[9]

$$\begin{aligned} B_{ik} &\leftarrow B_{ik} \frac{\sum_j C_{kj} X_{ij} / (BC)_{ij}}{\sum_j C_{kj}}, \\ B_{ik} &\leftarrow \frac{B_{ik}}{\sum_l B_{lk}}, \\ C_{kj} &\leftarrow \sqrt{C_{kj} \sum_i B_{ik} X_{ij} / (BC)_{ij}}, \end{aligned} \quad (23)$$

由于采用了一些近似处理, 所以该算法中的正则化参数 α, β 没有出现在迭代规则中. 关于这一点, 还有待进一步的研究.

2.3 非负稀疏编码(NNSC)

根据稀疏编码原则^[10], Hoyer^[11]提出了一种非负稀疏编码(NNSC), 使分解后的系数具有比较好的稀疏特性. 也就是求解如下的约束优化问题:

$$\begin{aligned} \min L_{NNSC}(B, C) &= \sum_{ij} [X_{ij} - (BC)_{ij}]^2 + \lambda \sum_{kj} C_{kj} \\ \text{s.t.} \quad &B \geq 0, C \geq 0, \sum_i B_{ik}^2 = 1, \forall k. \end{aligned} \quad (24)$$

其迭代算法如下:

$$\begin{aligned} B &\leftarrow B - \eta(BC - X)C^T, \\ B_{ik} &\leftarrow 0, B_{ik} < 0, \\ B_{ik} &\leftarrow \frac{B_{ik}}{\sqrt{\sum_i B_{ik}^2}}, \\ C &\leftarrow C \cdot (B^T X) / (B^T BC + \lambda), \end{aligned} \quad (25)$$

其中 η 为学习速率. 该学习算法的一个缺点是基向量学习是加性迭代, 不能很好地保持非负特性, 对负值必须使其强制置零. Hoyer^[12]将其算法应用在视觉感知的建模研究中.

2.4 稀疏非负矩阵分解(SNMF)

同样根据稀疏编码原则, 我们提出了如下稀疏非负矩阵分解(SNMF)^[13]:

$$\begin{aligned} \min L_{\text{SNMF}}(B, C) = & \sum_{ij} \left(X_{ij} \log \frac{X_{ij}}{(BC)_{ij}} - X_{ij} + (BC)_{ij} \right) \\ & + \beta \sum_{kj} C_{kj} \\ \text{s.t. } & B \geq 0, C \geq 0, \sum_i B_{ik} = 1, \forall k. \end{aligned} \quad (26)$$

其乘性迭代规则如下:

$$\begin{aligned} B_{ik} & \leftarrow B_{ik} \sum_j C_{kj} X_{ij} / (BC)_{ij}, \\ B_{ik} & \leftarrow \frac{B_{ik}}{\sum_l B_{lk}}, \\ C_{kj} & \leftarrow \frac{C_{kj} \sum_i B_{ik} X_{ij} / (BC)_{ij}}{1 + \beta}. \end{aligned} \quad (27)$$

和 NNSC 的学习算法相比, 该算法全部采用乘性迭代规则, 能很好地保持数据的非负特性. 另外, 关于非负矩阵分解的稀疏性约束详细讨论见文献[14].

2.5 其他算法

从算法优化的角度来看, 非负矩阵分解是一种满足非负性约束下的优化问题. 除了上面提到的几个改进算法以外, 在非负性约束的基础上, 人们还提出了加权的(正)非负矩阵分解^[4,15], 多重线性引擎^[16]和正张量分解^[17]. Wang^[18]将Fisher约束引入到分解中, Ahn^[19]还将非负矩阵的单层神经网络表示推广到多层的情形.

3 非负矩阵分解的应用

由于算法实现的简便和有效性, 非负矩阵分解已成为模式识别研究领域特征提取和数据降维的一种新方法, 在高维数据处理中有着广泛的应用前景.

在图像处理领域, Lee和Seung^[20]将非负矩阵分解应用于机器人对外界环境感知的非监督学习中, 用以提取图像中有意义的部分特征. 由于非负性约束使得分解的基向量和组合系数中的大量元素为零或接近于零, 因此这种表示方式属于稀疏编码, 占用的存储空间少. 在有遮挡的情况下, 非负矩阵分解具有很强的鲁棒性能. Lee等^[21]将非负矩阵分解方法用于动态PET图像序列的分析, 实验表明采用非负矩阵分解可以得到和传统因子分析一样的结果, 并

且更有利于从图像序列中实现图像分割和特征提取. Ramanath^[22]还将非负矩阵分解用于彩色谱空间的分析.

Kawamo^[23]将非负矩阵分解用于音乐信号的分析, 其基本思想是将音乐的幅度谱(利用离散傅里叶变换得到)进行分解, 从而得到具有单个音调的特征数据. 一些研究结果表明非负矩阵分解也非常适合声音数据分析^[24]. Novak^[25]还将非负矩阵分解用于自然语言的处理.

在文本知识挖掘领域, Xu^[26]将非负矩阵分解用于文本类数据的分析. 对这类稀疏性特别强的数据来说, 非负矩阵分解更是优于传统的文本特征提取方法. 在生物信息学研究领域, Seppanen和Brunet^[27]将非负矩阵分解用于基因、分子等数据的分析.

下面介绍非负矩阵分解算法在网络安全入侵检测、灰度图像的数字水印技术和脑电信号处理应用的具体实例²⁾, 希望引起大家的重视.

3.1 非负矩阵分解在网络安全入侵检测的应用

入侵检测是对计算机和网络资源上的恶意使用行为进行识别和响应的处理过程^[28]. 入侵检测的根本任务就是从这些数据中去发现其中存在的误用或异常. 如果把入侵检测看成一种数据处理过程, 那么我们就可以采用数据挖掘技术来建立入侵检测系统.

从系统调用类别数目来讲, 在 UNIX 环境下, 大约有 200 种不同的系统调用. 对于每个序列, 系统调用出现的频率特征是一个维数为系统调用类别数目的向量. 于是, 降维是必要的. 针对这种频率特征的非负性, 我们选择非负矩阵分解来作特征提取, 用线性回归作检测.

这里采用的实验数据来自于 UNM 的合成 sendmail³⁾和实际的 named⁴⁾两组数据. 在实验中, 系统调用类别数量在 200 左右. 为了使用非负矩阵分解方法, 首先将序列作连续的划分, 得到一些短序列, 然后统计每个短序列中各个系统调用出现的次数, 得到一个 $X_{d \times m}$ 矩阵, 其中的 X_{ij} 表示第 i 个系统调用出现在第 j 个短序列中的次数. 短序列长度和其中含有的系统调用类别数目根据不同的数据集来确定.

1) Seppanen J K, Hollmén J E, Bingham E, et al. Nonnegative matrix factorization on gene expression data. Bioinformatics 2002, poster 49, Bergen, 2002

2) 刘维湘. 非负矩阵分解及其应用. 西安交通大学博士学位论文, 2005

3) <http://www.cs.unm.edu/~immsec/data/>

4) <http://www.cs.unm.edu/~immsec/data/live-named.html>

合成的 *sendmail* 数据在文献[29]中有详细描述, 它包含了 300 多万条系统调用, 其中含有 16 个非正常序列. 在这 16 个非正常序列中, 最长的序列含有 1861 个系统调用, 最短的含有 254 个系统调用. 我们选择长度为 1500 来对正常数据进行分段, 得到了 2250 个正常短序列. 因此, 最后得到的数据为 16 个非正常数据和 2250 个正常数据. 在这个数据集中, 出现的基本系统调用是 53 个. 但是, 这里根据 SunOS 映射文件 (该文件是用十进制数来表示不同的系统调用) 考虑 182 种系统调用, 最后统计各个系统调用在每个短序列中出现的次数, 得到一个 182×2266 矩阵, 其中前 2250 列数据为正常数据, 后 16 列为异常数据. 选用前 30 个正常数据作为训练数据, 剩下的 2220 个正常数

据和 16 个异常数据作为测试数据.

实际的 *named* 数据包含一个有 900 万个系统调用的正常序列和 2 个成功入侵的异常序列, 该数据在文献[30]有详细描述. 在这 2 个入侵序列中, 一个含有 969 个系统调用, 另一个含有 831 个系统调用. 使用长度为 1000 来对正常数据进行分段, 得到 9230 个正常短序列. 根据系统调用的映射文件, 共有 164 种系统调用. 最后, 我们得到一个 164×9232 的矩阵, 前 9230 列为正常数据, 后 2 列为异常数据. 选用前 100 个正常数据作为训练数据, 其余的 9130 个正常数据和 2 个异常数据作为测试数据(见表 1).

我们采用线性回归方法来进行检测, 表 2 和 3 分别给出了 PCA 和 NMF 两种方法在两类数据上的正检

表 1 实验数据集

数据集	系统调用种类数	训练用正常数据数	测试用正常数据数	异常数据数
sendmail	182	30	2220	16
named	164	100	9130	2

表 2 PCA 和 NMF 在 *sendmail* 数据集上的检测结果比较

程序	方法	r	ε	TDR/%	FDR/%
Sendmail	PCA	1	5×10^{-5} 1×10^{-5}	93.75 100	0.09 3.38
		2	5×10^{-5} 1×10^{-5}	81.25 93.75	0 3.06
		3	5×10^{-5} 1×10^{-5}	93.75 100	4.19 6.62
		10	5×10^{-5} 1×10^{-5}	93.75 100	7.43 7.75
	NMF	1	5×10^{-8} 1×10^{-8}	93.75 100	0.05 0.36
		2	5×10^{-8} 1×10^{-8}	100 100	0.41 0.41
		3	5×10^{-8} 1×10^{-8}	100 100	0.41 0.41
		10	5×10^{-8} 1×10^{-8}	100 100	0.41 0.41

表 3 PCA 和 NMF 在 *named* 数据集上的检测结果比较

程序	方法	r	ε	TDR/%	FDR/%
Named	PCA	1	2×10^{-4} 1×10^{-4}	100 100	0.28 0.38
		2	2×10^{-4} 1×10^{-4}	100 100	0.26 0.34
		3	2×10^{-4} 1×10^{-4}	100 100	0.25 0.34
		10	2×10^{-4} 1×10^{-4}	50 100	0.38 0.48
	NMF	1	2×10^{-7} 1×10^{-7}	100 100	0 0.01
		2	2×10^{-7} 1×10^{-7}	100 100	0.21 0.35
		3	2×10^{-7} 1×10^{-7}	100 100	0.03 0
		10	2×10^{-7} 1×10^{-7}	100 100	0 0.21

率(TDR)和误检率(FDR). 从表中可以看出, NMF 方法可以得到较高的正检率和较低的误检率.

3.2 灰度图像的数字水印研究

数字水印是指利用信号处理的方法在数字化多媒体数据中嵌入隐蔽的标记信息. 这种标记信息一般是不可见的, 只有通过专用的检测器或阅读器才能提取得到. 数字水印技术已成为信息隐藏的一个重要研究方向¹⁾.

图1为数字水印系统的一般模型. 通常, 我们把具有某种含义的水印信息(W_i , 如公司的标志图案或随机码等)通过一种编码技术嵌入到需要保护的载体数据(H , 如图像、视频等)中; 然后, 加入水印的载体数据在其散发传递过程中会受到各种各样的攻击(如在传输中的噪声、压缩处理带来的变化或有意的破解攻击等); 最后, 我们需要从受到攻击后的载体数据中解码出含有的水印信息(W'_i), 并通过一定的检测方法来比较前后水印信息的不同, 从而判断受保护数据的使用是否合法等.

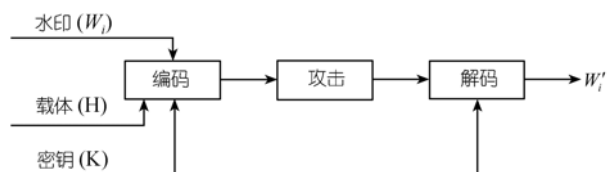


图1 数字水印系统的一般模型

按水印隐藏的空间划分, 可以将数字水印划分为时(空)域数字水印、频域数字水印、时/频域数字水印和时间/尺度域数字水印. 现在大多数的水印技术都是采用变换域方法, 只要构成一种信号变换, 就有可能在其变换空间上隐藏水印. 这也就是我们将非负矩阵分解应用在灰度图像的直接原因.

将非负矩阵分解看作一种信号变换, 在其系数矩阵中嵌入水印信息. 下面的实验采用全局的水印嵌入方法, 并与基于DCT的方法^[31]进行对比.

选用Lena灰度图像作实验原图, 该图像尺寸为 256×256 , 把像素值归一化到 $[0,1]$ 区间, 并用高斯随机序列(含有1000个数值)作为水印信息^[31]. 为了对比起见, 我们用1000个随机序列(其中第500个为嵌入的原始水印信息)和检测得到的水印作比较. 下面是考虑了不同攻击情况下的水印图像及其检测效果.

图2是在没有受到任何攻击下的水印图像及其检测结果. 从图中可以看出, 所得到的水印图像以及水印的检测反应值很接近.

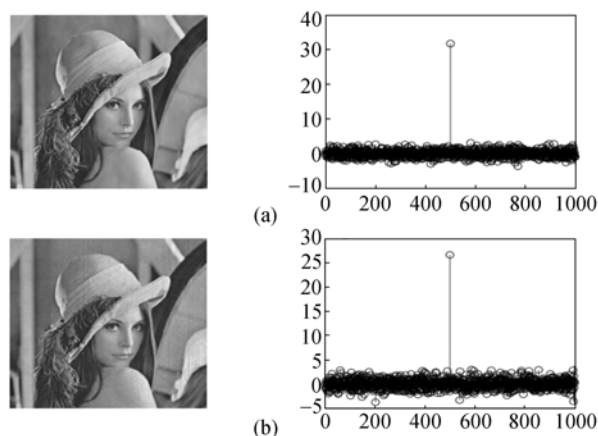


图2 没有攻击情况下的水印图像以及检测反应值对比
(a) DCT; (b) NMF

图3是经压缩处理的水印检测对比. 这里采用 8×8 分块DCT方法压缩, 每块只取低频部分的前10个系数. 由图3可以看出, 在这种情况下, 由NMF所得的水印检测值要低于采用DCT变换得到的水印检测值.

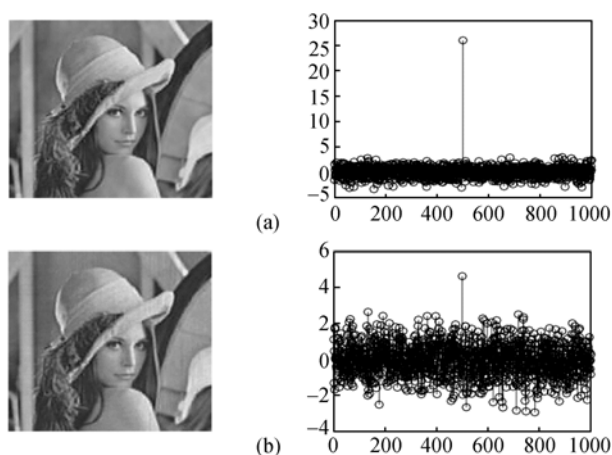


图3 经压缩处理的水印图像以及检测反应值对比
(a) DCT; (b) NMF

图4是在像素值发生尺度变化情况下的水印检测对比结果. 将加入水印信息的图像的像素值各乘以1.1得到受攻击的图像. 在此情形下, 由NMF所得的水印检测值要高于采用DCT变换得到的水印检测值.

1) <http://www.assuredigit.com/assure.asp?link=shuiyin>

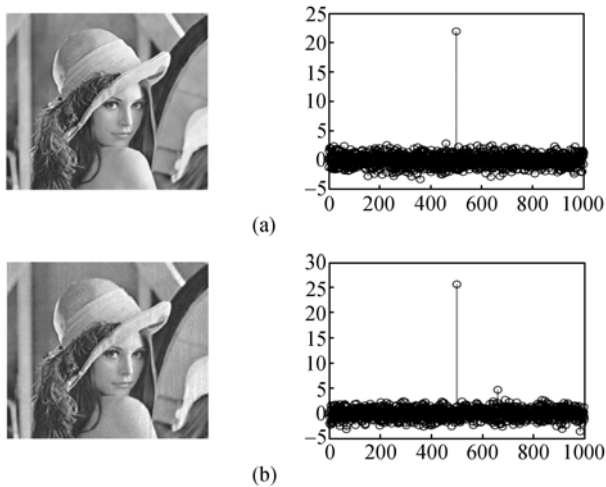


图 4 像素值发生尺度变化时的水印图像以及检测反应值对比
(a) DCT; (b) NMF

图 5 是在剪切情况下的水印检测对比, 将图像的左 100 列设置为 0 来实现剪切效果. 在这种情况下, 我们发现基于 NMF 的水印检测值远大于基于 DCT 的水印检测值, 这正是 NMF 基于部分表示特性所带来的好处.

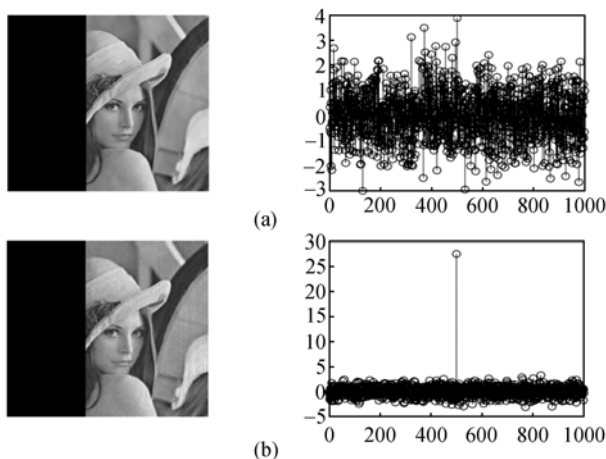


图 5 剪切情况下的水印图像以及检测反应值对比
(a) DCT; (b) NMF

3.3 脑电信号的特征提取和分类^[32]

大脑通常依赖于人体的外周神经和肌肉组织等与外部环境进行交流或控制外界设备. 人体在睡眠、兴奋或者接受外界信号刺激等不同状态下, 脑内电

活动会表现出不同的模式, 应用传感器可以把这种脑内电活动转化为脑电信号. 基于脑电信号的脑机接口(brain computer interface, BCI)研究可以帮助具有严重神经或肌肉伤残的人与外界进行沟通. 我们这里讨论自发电位的脑电信号^[33].

一般的自发脑电信号的频率范围是在 1~20 Hz, 更宽的在 1~50 Hz. 受非负矩阵分解在音乐信号处理^[23]应用的启发, 将非负矩阵分解应用在脑电信号的幅度谱分解中, 进行特征提取和分类.

这里实验的数据¹⁾来源于Keirn等人^{[34]2)}的工作. 在七位受试人员头部的不同位置获取 6 个通道的信号, 信号的采样频率是 250 Hz, 持续时间是 10 s, 然后经过一组带通模拟滤波器(其频率带宽是 0.1 到 100 Hz)输出. 文献[35,36]给出了关于该数据的描述和相关的研究应用.

按照下面的规则来截取每个通道的数据段: 每个数据段长度为 L , 相邻的两个数据段之间的重合长度选为 $OL = L/2$. 为了方便进行快速傅里叶变换, 选取 $L = 2^k$. 将得到的每个数据段视为一个样本, 其维数就为 L . 再经过快速傅里叶变换后, 取其幅度谱中的前 $(L/2+1)$ 就得到一个维数为 $(L/2+1)$ 的样本. 考虑 6 个通道情况, 则每个样本(幅度谱)最后的维数将是 $6 \times (L/2+1)$. 如果选取 $k=6$, 则每个样本(幅度谱)的维数高达 198. 下面用非负矩阵分解来进行降维, 然后采用线性判别分类器^[34]来实现分类.

实验中比对的子空间维数从 1 增加到 70, 迭代次数均为 300. 鉴于非负矩阵分解初始化的影响, 将实验重复进行 5 次.

实验中每位受试人员完成两个意识任务: 数学乘积运算(math)和写信(letter). 为了和文献[35]中的实验结果进行对比, 每个数据段窗口长度为 $L = 2^6 = 64$, 各个数据段之间的覆盖长度均为 $OL = 32$. 首先对第一天的试验数据进行分类. 将其中第一次试验数据作为训练样本, 其余的四次试验数据作为测试样本, 则每类训练样本数均为 77, 每类测试样本数均为 308. 我们对比了在不同子空间维数下两种不同任务的正确识别率和平均识别率. 表 4 给出了每位受试人员的数据在 5 次分类结果中的最高平均识别率及相应的单个任务识别率.

1) <http://www.cs.colostate.edu/eeg/index.html#Data>

2) Keirn Z A. Alternative modes of communication between man and machine. Masters thesis, Electrical Engineering, Purdue University, 1988

表4 各受试人员在同一天试验数据的两类分类结果(%)

受试人员	数学	写信	平均
1	98.7	97.7	98.2
2	53.2	95.5	74.4
3	38.3	96.4	67.4
4	64.9	94.2	79.5
5	50.6	68.5	59.6
6	74.4	79.2	76.8
7	99.4	93.8	96.6

从表4可以看出,最后的分类结果因人而异.这和每位受试人员在实验过程中的配合程度,对意识任务的控制以及信号的采集状况都有关系.实验结果表明:在5次的分类结果中,受试人员1的最高平均分类识别率可以达到98.2%,该结果好于文献[35]中的结果90%;受试人员7的分类结果也高达96.6%.

根据上面的实验结果,考虑受试人员1的试验数据,将受试人员1第一天的第一次试验数据作为训练样本,第二天的全部5次试验数据作为测试样本.因此,每类测试样本数均为385.实验结果中,最高的平均分类识别率可以达到80.6%,该结果好于文献[35]中的结果75%.

为了与Garrett^[36]的实验结果进行对比,每个数据段窗口长度为 $L=125$,各个数据段之间的覆盖长度均为 $OL=63$.考虑在第一天的试验数据上进行分类,将其中第一次试验数据作为训练样本,其余的四次试验数据作为测试样本.每类训练样本数均为39,每类测试样本数均为156.

表5给出了应用NMF方法的实验结果,其中,a,b,c,d,e分别是对应于5个意识任务的最高识别率结果,f为平均识别率最高时的结果.Garrett利用6阶自回归模型来表示信号,分别利用线性判别分析(LDA)、神经网络(NN)和支持向量机(SVM)来进行分类^[36],其结果如表6.通过对比可以看出,和线性判别分析分类所得结果相比,采用NMF方法提高了识

表5 应用NMF的多类分类结果(%)

	休息	数学	写信	旋转	计数	平均
a	75.6	84.6	84.6	52.6	46.2	68.7
b	67.9	94.9	61.5	48.7	48.7	64.4
c	58.3	90.4	94.9	51.9	51.3	69.4
d	63.5	48.1	28.2	63.5	32.1	47.1
e	58.3	82.7	92.3	43.6	68.6	69.1
f	64.1	90.4	91.0	53.8	51.3	70.1

表6 Garrett等的实验结果

分类器	休息	数学	写信	旋转	计数	平均
LDA(%)	47.3	45.1	51.1	38.8	44.5	44.8
NN(%)	64.3	47.3	54.7	51.1	47.3	52.8
SVM(%)	59.4	44.5	52.7	57.0	47.9	52.3

别率.根据最后的平均识别率,采用NMF方法还优于由神经网络和支持向量机得到的识别率.

另外,在实验结果中,分类识别率的稳定性不好,其原因可能与非负矩阵分解的初始化有关.

4 讨论

4.1 非负矩阵分解与认知

人脑是最有效的生物智能系统,它具有感知、识别、学习、联想、记忆、推理等功能.研究人脑并用机器来模拟这些功能,一直是信息科学发展中最有意义和极具挑战性的重大问题.在日常生活中,人们有80%以上的信息来源于视觉.然而,关于人的视觉感知机理我们却知之甚少.

在知觉研究200多年的历史中,关于知觉过程有“整体论”和“原子论”之争¹⁾.前者认为人对外部世界的感知是全局性结构感知先于局部性结构感知,即知觉过程始于对物体的整体性感知,是从大范围性质到局部性质,视觉信息处理中由粗到细的多分辨率分析就是这一观点的实例;而后者认为知觉过程始于对物体的特征性质或简单组成部分的感知与分析,是由局部性质到大范围性质.另外,一些视觉感知过程的实验又支持一种融合观点:全局性和局部性感知是互动的,小尺度和大尺度感知是并行的、相互作用的;同时,视觉感知具有小范围竞争、大范围协作的特点.然而,无论是从整体到局部,或者是从局部到整体,或两者同时兼有,弄清局部特征和整体特征之间的关系具有重要意义.非负矩阵分解为分析局部特征和整体特征之间的关系提供了一种思路,即整体特征是局部特征的非负线性组合,局部特征在构成整体特征时不会产生正负抵消的情况.这种非负分解特性更合乎人类视觉感知的直观体验.

认知科学领域的重要研究内容之一是探索人脑信息的处理过程,包括知觉、注意、记忆、动作、语言、推理、思考、意识乃至情感在内的各个层面的认知活动,研究认知过程的信息处理就是要用“计算”来“表达”人类智能²⁾.非负矩阵分解为我们提供了一

1) 陈霖. 知觉过程是从哪里开始的? 中国科学院第十二次院士大会学术报告. <http://www.casad.ac.cn/2005-4/2005418101856.htm>

2) 郑南宁. 认知与计算机视觉. 国家杰出青年科学基金实施十周年学术报告, 2004

种用“计算”来“表达”智能的有效的数学分析方法。由于非负矩阵分解具有基于部分表示的能力,它与主成分分析方法相比,即使在物体有遮挡或图像被剪切的情况下,仍能够取得较高的识别率和检测率^[91],这也和人们能够根据部分的、不完整的信息进行推理的能力相吻合。

对图像来说,尽管非负矩阵分解能够从中得到有关物体基于部分的信息,然而非负性约束并没有明确考虑图像中的邻域结构信息。另外,大脑的感知由于存在多水平、多层次的协同整合,对复杂的环境具有很强的自适应能力。而非负矩阵分解还是一种单一水平的处理方法,如何把非负矩阵分解思想应用于多层次的感知过程还有待进一步的研究^[19]。

4.2 关于非负性的讨论

如文献[3]所述,非负矩阵分解中的非负性约束有着明确的物理意义,并且非负性在一定程度上导致了稀疏性。而传统主成分分析是将信号进行正交分解,其分解的结果中含有负值,缺失了实际的物理含义;独立成分分析则假定各源分量满足统计独立的条件。有研究表明,从基选择的角度看,稀疏性要优于统计独立条件要求^[37]。关于非负性在统计信号中的应用,可以参考文献[38]。很多的信号分解方法,例如傅里叶分析,以及小波分析,要求其基信号是正交的,甚至是完备的。进一步探讨信号处理中的正交性、独立性、非负性以及稀疏性的特点是很有意义的。

4.3 非负矩阵分解中有待进一步研究的问题

非负矩阵分解已有一些成功的应用,但在理论和实践方面还需要进一步深入研究和探讨:

(1) 如何最小化 $\sum_{i,j} |X_{ij} - (BC)_{ij}|$? 非负矩阵分解

作为一个约束优化问题,首先需要寻找对其优化的目标函数和相关的约束。

(2) 寻找新的快速迭代计算方法。对于大的数据集来说,文献[5]提出的两个乘性迭代算法依然有其局限性。因此,需要寻找新的快速迭代计算方法^{[39]2)}。

(3) 在分解过程中,如何确定低维特征数 r 的大小?

(4) 如何进行非负矩阵分解的初始化? 非负矩阵分解的结果往往和初始化有关,得到的是局部最优解(关于迭代规则的初始值选择,有研究者提出采用球形聚类方法来确定^[40])。

(5) 研究非负矩阵分解的解的特性。非负矩阵分解是一个逆问题,是不适定的。作为一个逆问题,需要进一步从理论上对非负矩阵分解的解的特性(惟一性、存在性和稳定性)进行研究^{3,4)}。

(6) 寻求将非负性约束和其他方法相结合的新的分解方法,如非负独立成分分析^[41]和非负Boltzmann机^[42]。

(7) 探讨其他基于部分表示的学习方法^[43,44]。

5 结论

非负矩阵分解是一种专门针对非负数据分析的矩阵分解方法,其基本思想是对非负数据进行线性非负分解。在功能上,它实现了对大脑的基于部分感知功能的模拟;在算法上,采用简单有效的乘性迭代规则,很好地保持了非负性;在应用中,非负性的数据大量存在,且非负分解的结果具有明确的物理含义;作为一种低秩逼近算法,能有效节约存储和计算资源。非负矩阵分解算法在网络安全入侵检测、灰度图像的数字水印技术和脑电信号处理方面的应用,证明非负矩阵分解在智能信息处理和模式识别的大规模数据处理与分析中有着十分重要的应用意义。

从已有的对非负矩阵分解的改进算法以及上面的讨论来看,需要在理论上对非负矩阵分解做更深入的研究,寻求新的分解算法,并进一步探讨该方法在智能信息处理和模式识别领域中的应用。

另外,还可以参考本文列出的其他有关非负矩阵分解方法的综述和讨论^{[45]5)}。

致谢 感谢所有匿名审阅者提供的建设性意见。本工作得到国家自然科学基金委员会创新研究群体科学基金研究计划(批准号:60021302)和国家自然科学基金项目(批准号:60205001)的资助。

参 考 文 献

- 1) Hubert L J, Meulman J J, Heiser W J. Two purposes for matrix factorization: A historical appraisal. *SIAM Review*, 2000, 42: 68~82[DOI]

1) 同244页脚注2)
2) Liu W G, Yi G L. Existing and New Algorithms for Nonnegative Matrix Factorization. <http://www.cs.utexas.edu/users/liuwg/383CProject/CS383CProject.htm>
3) Donoho D, Stodden V. When does non-negative Matrix factorization give a correct decomposition into parts? Tech Report, Department of Statistics, Stanford University, 2003
4) Chu M, Diele F, Plemmons R, et al. Optimality, computation, and interpretations of nonnegative (Non-negative) matrix factorizations. Submitted to the *SIAM Journal on Matrix Analysis*, 2004
5) http://www.wfu.edu/~plemmons/papers/chu_ple_survey.pdf, <http://www.ece.utexas.edu/~bevans/courses/ee381k/projects/spring03/tropp/LitSurveyReport.pdf>

- 2 Duda R O, Hart P E, Stork D G. Pattern Classification. Second edition. New York: John Wiley & Sons, 2001
- 3 Lee D D, Seung H S. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 1999, 401: 788~791[DOI]
- 4 Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 1994, 5: 111~126
- 5 Lee D D, Seung H S. Algorithms for nonnegative matrix factorization. In: Leen T, Dietterich T, Tresp V, eds. *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press, 2000
- 6 Sajda P, Du S, Parra L, et al. Recovery of constituent spectra using non-negative matrix factorization. *Proc SPIE*, 2003, 5207: 321~331[DOI]
- 7 Cichocki A, Amari S. Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications. New York: John Wiley & Sons, 2002
- 8 Liu W X, Zheng N N. Learning sparse features for classification by mixture models. *Pattern Recognition Letters*, 2004, 25(2): 155~161[DOI]
- 9 Li S Z, Hou X W, Zhang H J, et al. Learning spatially localized, parts-based representation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, 1: 207~212
- 10 Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 1996, 381: 607~609[DOI]
- 11 Hoyer P O. Non-negative sparse coding. *Proc Neural Networks for Signal Processing*, 2002, 557~565
- 12 Hoyer P O. Modeling receptive fields with non-negative sparse coding. *Neurocomputing*, 2003, 52-54: 547~552
- 13 Liu W X, Zheng N N, Lu X F. Non-negative matrix factorization for visual coding. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003, 3: 293~296
- 14 Hoyer P O. Non-negative matrix factorization with sparseness constraints. *Journal of Machine learning research*, 2004, 5: 1457~1469
- 15 Guillaumet D, Vitri  J, Schiele B. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recognition Letters*, 2003, 24(14): 2447~2454[DOI]
- 16 Paatero P. Least squares formulation of robust non-negative factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 1997, 37(1): 23~35[DOI]
- 17 Welling M, Weber M. Positive tensor factorization. *Pattern Recognition Letters*, 2001, 22(12): 1255~1261[DOI]
- 18 Wang Y, Jia Y, Hu C, et al. Fisher non-negative matrix factorization for learning local features. *Asian Conference on Computer Vision*, Korea, 2004
- 19 Ahn J H, Choi S, Oh J H. A multiplicative up-propagation algorithm. *Proc Internal Conference on Machine Learning*, 2004: 17~24
- 20 Lee D D, Seung H S. Learning in intelligent embedded systems. *Proceedings of the Embedded Systems Workshop*, Cambridge, Massachusetts, USA, 1999, http://hebb.mit.edu/people/seung/papers/lee_es99.pdf
- 21 Lee J S, Lee D D, Choi S, et al. Non-negative matrix factorization of dynamic images in nuclear medicine. *Proceedings of IEEE Nuclear Science Symposium Conference Record*, San Diego, California, 2001, 4: 2027~2030
- 22 Ramanath R, Kuehni R G, Snyder W E, et al. Spectral spaces and color Spaces. *Color Research and Application*, 2004, 29(1): 29~37[DOI]
- 23 Kawamoto T, Hotta K, Mishima T, et al. Estimation of single tones from chord sounds using non-negative matrix factorization. *Neural Network World*, 2000, 3: 429~436
- 24 Cho Y C, Choi S, Bang S Y. Non-negative component parts of sound for classification. *Proc IEEE Int'l Symp Signal Processing and Information Technology*, Darmstadt, Germany, 2003
- 25 Novak M, Mammone R. Improvement of non-negative matrix factorization based language model using exponential models. *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, 190~193
- 26 Xu W, Liu X, Gong Y. Document-clustering based on non-negative matrix factorization. *Proceedings of SIGIR'03*, Toronto, CA, 2003, 267~273
- 27 Brunet J P, Tamayo P, Golub T R, et al. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 2004, 101: 4164~4169[DOI]
- 28 戴英侠, 连一峰, 王航. 系统安全和入侵检测. 北京: 清华大学出版社, 2002
- 29 Forrest S, Hofmeyr S A, Somayaji A, et al. A sense of self for unix processes. *Proceedings of IEEE Symposium on Computer Security and Privacy*, 1996, 120~128
- 30 Warrender C, Forrest S, Pearlmuter B. Detecting intrusions using system calls: Alternative data models. *Proceedings of IEEE Symposium on Security and Privacy*, 1999, 133~145
- 31 Cox I J, Kilian J, Leighton T, et al. Secure spread spectrum watermarking for multimedia. *IEEE Trans on Image Processing*, 1997, 6(12): 1673~1687[DOI]
- 32 Liu W X, Zheng N N, Li X. Nonnegative Matrix Factorization for EEG Signal Classification. *ISNN* (2), 2004: 470~475
- 33 Wolpaw J R, Birbaumer N, McFarland D J, et al. Brain computer interfaces for communication and control. *Clinical Neurophysiology*, 2002, 113(6): 767~791[DOI]
- 34 Keirn Z A, Aunon J I. A new model of communication between man and his surroundings. *IEEE Transactions on Biomedical Engineering*, 1990, 37: 1209~1214[DOI]
- 35 Anderson C W, Kirby M. EEG subspace representations and feature selection for brain computer interfaces. *Proceedings of the 1st IEEE Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction*, 2003, 475~483
- 36 Garrett G, Peterson D A, Anderson C W, et al. Comparison of linear and nonlinear methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitative Engineering*, 2003, 11(2): 141~144[DOI]
- 37 Saito N, Larson B M, Benichou B. Sparsity vs. statistical independence from a best-basis viewpoint. In: Aldroubi A, Laine A F, Unser M A, eds. *Proc SPIE Wavelet Applications in Signal and Image Processing VIII*, 2000, 4119: 474~486
- 38 Saul L K, Sha F, Lee D D. Statistical signal processing with non-negativity constraints. *Proceedings of the Eighth European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, 2: 1001~1004
- 39 Liu W X, Zheng N N, Li X. Relative gradient speeding up additive updates for nonnegative matrix factorization. *Neurocomputing*, 2004, 57: 493~499[DOI]
- 40 Wild S M, Curry J, Dougherty A. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 2004, 37: 2217~2232[DOI]
- 41 Plumbley M D. Conditions for non-negative independent component analysis. *IEEE Signal Processing Letters*, 2002, 9(6): 177~180[DOI]
- 42 Downs O B, MacKay D, Lee D D. The nonnegative boltzmann machine. In: Solla S A, Leen T K, Muller K R, eds. *Advances in Neural Information Processing Systems 12*, 2000, 428~434
- 43 Ross D, Zemel R. Multiple cause vector quantization. In: Becker S, Thrun S, Obermayer K, eds. *Advances in Neural Information Processing Systems 15*. Cambridge: MIT Press, 2003
- 44 Ge X J, Iwata S. Learning the parts of objects by auto-association. *Neural Networks*, 2002, 15: 285~295[DOI]
- 45 汪鹏. 非负矩阵分解: 数学的奇妙力量. *计算机教育*, 2004, 10: 38~40

(2005-07-13 收稿, 2005-11-02 接受)