

非负矩阵分解：模型、算法和应用*

章祥荪¹, 张忠元²

(1. 中国科学院 数学与系统科学研究院, 北京 100190; 2. 中央财经大学 统计与数学学院, 北京 100081)

摘要:近年来,非负矩阵分解模型已经成为数据挖掘领域中最成功的模型之一。该模型能够自动从一组高维向量中提取隐含模式,从而被广泛应用于降维、无监督学习(图像处理、聚类和双聚类等)和预测当中。本文将从它的发展历史、数学表达形式、算法和热点应用等几个层面对非负矩阵分解模型进行综述。简言之,该模型具有较好的可解释性,模型简单,易于理解操作,可拓展性强,该模型和无监督学习领域中其它被广泛采用的模型关系紧密,且有广泛的应用空间,数值表现优异。同时作为一项新兴技术,该模型亦有许多有趣的问题值得进一步深入研究。

关键词:非负矩阵分解;乘性迭代算法;K-means;潜在语义分析;图像处理;数据聚类;社团结构探测

中图分类号:O151.21;TP391

文献标志码:A

文章编号:1672-6693(2013)06-0001-08

1 非负矩阵分解的研究历史

非负矩阵分解(Nonnegative matrix factorization, NMF)^[1-3]起源于主成分分析(Principal component analysis, PCA)^[4-5]。PCA最早可以追溯到1901年,被Karl Pearson^[6]最先提出,稍后在上世纪30年代又由Harold Hotelling独立重新提出并被命名^[7-8]。PCA是数据降维的一个基本技术,它在最小二乘的意义下能够计算得到一组最优基向量来近似高维数据。基向量的数目比原始数据的维数要少得多,因此将原始数据表达为这些基向量的线性组合就会将其变换到一个低维空间中。这种降维可以用来改进数据分析算法的处理能力,也可以用于数据的特征选择和特征提取。

主成分分析得到的基向量中的分量有正值也有负值,而数据被表达成这些基向量的线性组合,组合系数中也是有正有负。PCA的最优近似性质就可以被认为是这种可约减性质导致的一个结果。然而在许多应用中,负值是和客观现实相矛盾的。举例来说,图像中的像素如果是负值,就很难对其进行合理而直观的解释。并且在实际应用中,比如复杂网络分析、图像处理、生物和文本分析中,非负数据分析是很普遍的,为了更好地分析非负数据,同时提高结果的可解释性,一些研究者建议寻找非负基向量来表达原始数据。非负矩阵分解模型由此被提出。事实上,该模型可以追溯到上世纪70年代^[9],在上世纪90年代Paatero^[3]和Lee & Seung^[1-2]的工作之后引起了广泛的注意。严格地说,这个想法可以被表达成:将 $n \times m$ 维的非负矩阵 \mathbf{X} 分解为2个非负因子矩阵 \mathbf{F} 和 \mathbf{G} 使得 $\mathbf{X} \approx \mathbf{F}\mathbf{G}^T$,其中 \mathbf{F} 和 \mathbf{G} 的维数分别是 $n \times k$ 和 $m \times k$ 的, k 是预先给定的参数。这个模型就被称为非负矩阵分解模型。假设 $k \ll n, m$,因此只有当恰当地找到了数据本身的类结构的时候, $\mathbf{F}\mathbf{G}^T$ 才会尽可能地拟合 \mathbf{X} ^[10]。

近年来,关于非负矩阵分解的研究从未间断,该模型的理论性质不断被发现,而其应用也不断得到拓展。它已成为线性代数^[11]、优化^[12]、复杂网络分析^[13-18]、模式识别^[19]、图像处理^[1]、多媒体数据分析^[20]、文本挖掘^[21-24]和生物信息学^[25-28]等多个领域的一个热点,成功应用于特征提取和降维、聚类、半监督聚类、双聚类和协同过滤等问题。

这一节最后给出一个NMF区别于主成分分析的理论性质:界值性质^[29-30]。就像前面提到的,NMF区别于

* 收稿日期:2013-08-07 网络出版时间:2013-11-20 14:46

资助项目:国家自然科学基金(No. 11131009; No. 61203295);中央财经大学科研创新团队支持计划(2013)

作者简介:章祥荪,男,研究员,博士生导师,研究方向为最优化理论与应用、生物信息学, E-mail: zxs@amt.ac.cn; 通讯作者:张忠元,

E-mail: zhyuanzh@gmail.com

网络出版地址: http://www.cnki.net/kcms/detail/50.1165.N.20131120.1446.201306.1_028.html

PCA 最重要的一点就是其非负性的约束,那么这种非负性约束的本质内涵是什么呢?下面的界值定理部分地回答了这个问题。

定理 1 (界值性质) 令 $0 \leq \mathbf{X} \leq M$ 是目标矩阵,其中 M 是固定的正常数。 \mathbf{F}, \mathbf{G} 是非负矩阵,满足于 $\mathbf{X} = \mathbf{F}\mathbf{G}^T$ 。则存在一个对角矩阵 $\mathbf{X} \geq 0$ 使得 $\mathbf{X} = \mathbf{F}\mathbf{D}(\mathbf{G}\mathbf{D}^{-1})^T = \mathbf{F}^* \mathbf{G}^{*T}$, 且 $0 \leq \mathbf{F}_{ij}^* \leq \sqrt{M}, 0 \leq \mathbf{G}_{ij}^* \leq \sqrt{M}$ 。特别地,若 \mathbf{X} 是对称的,且 $\mathbf{F} = \mathbf{G}^T$, 则 $\mathbf{G}^* = \mathbf{G}$ 。

注 $0 \leq \mathbf{X} \leq M$ 表示 $0 \leq X_{ij} \leq M, i=1, 2, \dots, n, j=1, 2, \dots, m$ 。

本文剩余部分组织如下:第 2 部分介绍 NMF 的乘性迭代算法(Multiplicative update rules),第 3 部分讨论了 NMF 模型和其它模型的关系,第 4 部分介绍了 NMF 的几个经典应用,并以社团结构探测问题为例介绍 NMF 在复杂网络分析中的应用,最后第 5 部分给出本文的总结和展望。

2 非负矩阵分解的乘性迭代算法

从最优化的角度,NMF 模型可以被表述成如下的非线性规划模型

$$\begin{aligned} \min J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T) \\ \text{s. t. } \mathbf{F} \geq 0, \mathbf{G} \geq 0 \end{aligned}$$

其中 $J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T)$ 是度量 \mathbf{X} 和 $\mathbf{F}\mathbf{G}^T$ 之间差异的费用函数。

可以看到,矩阵 \mathbf{F} 和 \mathbf{G} 中的所有元素都是最优化的变量,因此这是一个大规模优化问题,而且在实践中, $J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T)$ 所取的形式具有这样的性质:在固定 \mathbf{G} 的时候,对 \mathbf{F} 是凸的,在固定 \mathbf{F} 的时候,对 \mathbf{G} 是凸的。因此,绝大多数求解 NMF 的算法都是迭代地轮换地最优化 \mathbf{F} 和 \mathbf{G} 。计算 NMF 的一般算法可以被写成算法 1,不过可以采用不同的方法实现第 2 行和第 3 行,比如轮换最小二乘法(Alternative least squares)^[3]、投影梯度算法^[31]和字典学习算法^[14,32]等。而在实际计算中使用最广泛的是乘性迭代算法(Multiplicative update rules)^[1-2]。

算法 1 非负矩阵分解(NMF)一般算法框架

输入: $\mathbf{F}^{(0)}, \mathbf{G}^{(0)}, t=1$

输出: \mathbf{F}, \mathbf{G} 。

1: while 1 do

2: 固定 $\mathbf{G}^{(t-1)}$, 计算 $\mathbf{F}^{(t)}$ 使得 $J(\mathbf{X}, \mathbf{F}^{(t)} \mathbf{G}^{(t-1)T}) \leq J(\mathbf{X}, \mathbf{F}^{(t-1)} \mathbf{G}^{(t-1)T})$;

3: 固定 $\mathbf{F}^{(t)}$, 计算 $\mathbf{G}^{(t)}$ 使得 $J(\mathbf{X}, \mathbf{F}^{(t)} \mathbf{G}^{(t)T}) \leq J(\mathbf{X}, \mathbf{F}^{(t)} \mathbf{G}^{(t-1)T})$;

4: 检验算法是否终止;

5: if 算法终止性条件被满足 then

6: $\mathbf{F} = \mathbf{F}^{(t)}$

7: $\mathbf{G} = \mathbf{G}^{(t)}$

8: 退出

9: end if

10: $t = t + 1$

11: end while

费用函数可以有不同的选择,最经常被选用的是 $J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T) = \|\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2$ 或者是广义的 Kullback-Leibler divergence $J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T) = \sum_{i,j} (X_{ij} \log X_{ij} / (\mathbf{F}\mathbf{G}^T)_{ij} - X_{ij} + (\mathbf{F}\mathbf{G}^T)_{ij})$ 。在不同目标函数下 \mathbf{F} 和 \mathbf{G} 的调整规则总结在表 1 中^[33]。

所有这些公式实际上都是由非线性规划中的梯度下降算法导出的。以 $J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T) = \|\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2$ 为例来说明算法的推导过程。

首先,费用函数 $J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T) = \|\mathbf{X} - \mathbf{F}\mathbf{G}^T\|_F^2$ 对 \mathbf{F} 的偏导数是 $\frac{\partial}{\partial \mathbf{F}_{au}} J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T) = - \sum_j (X_{aj} - \sum_t (F_{at} G_{jt})) G_{ju}$ 。取步长为 $\eta_{au} = F_{au} / (\mathbf{F}\mathbf{G}^T \mathbf{G})_{au}$, 则

$$\mathbf{F}_{au} = \mathbf{F}_{au} - \eta_{au} \frac{\partial}{\partial \mathbf{F}_{au}} J(\mathbf{X} \parallel \mathbf{F}\mathbf{G}^T) = \mathbf{F}_{au} \frac{(\mathbf{X}\mathbf{G})_{au}}{(\mathbf{F}\mathbf{G}^T \mathbf{G})_{au}}$$

通过转换 \mathbf{F} 和 \mathbf{G} 的角色,可以容易地得到 \mathbf{G} 的调整公式。可以看到,算法并不是精确的梯度下降算法:步长是通过事先指定得到的,并不是通过线性搜索得到的。这样就需要证明算法的收敛性,即证明算法是单调下降的,并最终达到一个局部最优点。该算法的收敛性证明框架由文献[2]给出,这一框架源于统计中用于参数估计的 Expectation maximization 算法(EM)的收敛性证明,它已经成为类非负矩阵分解算法收敛性证明的一个一般框架。证明是基于辅助函数的。但是该证明仅可以保证算法是单调下降的^[34]。

文献[33]对这些目标函数和相应的迭代准则进行了系统比较,使用了 6 种人工数据,11 组生物基因组数据,2 组人脸数据和 22 组信号数据进行比较,结果表明传统的最小二乘目标函数和 KL divergence 目标函数的数值结果最好,正应了那句名言:大繁至简。

文献[33]对这些目标函数和相应的迭代准则进行了系统比较,使用了 6 种人工数据,11 组生物基因组数据,2 组人脸数据和 22 组信号数据进行比较,结果表明传统的最小二乘目标函数和 KL divergence 目标函数的数值结果最好,正应了那句名言:大繁至简。

表 1 不同的费用函数和相应的乘性迭代准则总结

Tab.1 Summary of different divergence functions and the corresponding multiplicative upalate rules

费用函数	乘性迭代算法	收敛法	费用函数	乘性迭代算法	收敛法
Csiszar's φ Divergence	—		Itakura-Saito(IS) Divergence	$\mathbf{F}_{ik} := \mathbf{F}_{ik} \frac{\sum_j \frac{\mathbf{X}_{ij}}{(\mathbf{F}\mathbf{G}^T)_{ij}^2} \mathbf{G}_{jk}}{\sum_j \frac{\mathbf{G}_{jk}}{(\mathbf{F}\mathbf{G}^T)_{ij}}}$	未证明
α -Divergence	$\mathbf{F}_{ik} := \mathbf{F}_{ik} \left(\frac{\left(\left(\frac{\mathbf{X}}{\mathbf{F}\mathbf{G}^T} \right)^{\alpha} \mathbf{G} \right)_{ik}}{\sum_l \mathbf{G}_{lk}} \right)^{\frac{1}{\alpha}}$	收敛		$\mathbf{G}_{jk} := \mathbf{G}_{jk} \frac{\sum_j \frac{\mathbf{X}_{ji}}{(\mathbf{F}\mathbf{G}^T)_{ji}^2} \mathbf{F}_{jk}}{\sum_j \frac{\mathbf{F}_{jk}}{(\mathbf{F}\mathbf{G}^T)_{ji}}}$	
Bregman Divergence	$\mathbf{F}_{ik} := \mathbf{F}_{ik} \frac{\sum_j \nabla^2 \varphi(\mathbf{F}\mathbf{G}^T)_{ij} \mathbf{X}_{ij} \mathbf{G}_{jk}}{\sum_j \nabla^2 \varphi(\mathbf{F}\mathbf{G}^T)_{ij} (\mathbf{F}\mathbf{G}^T)_{ij} \mathbf{G}_{jk}}$	收敛	K-L Divergence	$\mathbf{F}_{ik} := \frac{\mathbf{F}_{ik}}{\sum_j \mathbf{G}_{jk}} \sum_j \frac{\mathbf{X}_{ij}}{(\mathbf{F}\mathbf{G}^T)_{ij}} \mathbf{G}_{jk}$	收敛
β -Divergence	$\mathbf{G}_{jk} := \mathbf{G}_{jk} \frac{\sum_j \nabla^2 \varphi(\mathbf{G}\mathbf{F}^T)_{ij} \mathbf{X}_{ji} \mathbf{F}_{jk}}{\sum_j \nabla^2 \varphi(\mathbf{G}\mathbf{F}^T)_{ij} (\mathbf{F}\mathbf{G}^T)_{ji} \mathbf{F}_{jk}}$		Least Squares Error	$\mathbf{F}_{ik} := \mathbf{F}_{ik} \frac{(\mathbf{X}\mathbf{G})_{ik}}{(\mathbf{F}\mathbf{G}^T \mathbf{G})_{ik}}$	收敛
	$\mathbf{F}_{ik} := \mathbf{F}_{ik} \frac{\sum_j ((\mathbf{F}\mathbf{G}^T)_{ij}^{\beta-1} \mathbf{X}_{ij}) \mathbf{G}_{jk}}{\sum_j (\mathbf{F}\mathbf{G}^T)_{ij}^{\beta} \mathbf{G}_{jk}}$	收敛, 当 $0 \leq \beta \leq 1$ 时		$\mathbf{G}_{jk} := \mathbf{G}_{jk} \frac{(\mathbf{X}^T \mathbf{F})_{jk}}{(\mathbf{G}\mathbf{F}^T \mathbf{F})_{jk}}$	

注: 表 1 中的收敛性仅仅意味着算法已被证明是单调下降的, 并不一定收敛到局部最小点。

3 非负矩阵分解模型和其它模型的关系

分析 NMF 和其它模型的关系, 比如和 K-means、潜在语义分析(Probabilistic latent semantic indexing, PLSI)之间的关系, 可以有利于更好地理解它良好的数值表现和分析它的局限。基本结果如下。

1) 软 K-means 模型可以被写成对称 NMF 模型, 因而 K-means 可以认为是和 NMF 等价的^[35]。这就从理论上说明了 NMF 为什么适用于数据聚类。注意这个等价性的结论并不意味着 NMF 和 K-means 一定输出同样的结果, 因为它们采用完全不同的算法。

定理 2 $\min_{H \geq 0} \| \mathbf{W} - \mathbf{H}\mathbf{H}^T \|$ 和 \mathbf{X} 的软 K-means 聚类模型是等价的, 其中 $\mathbf{W} = \mathbf{X}^T \mathbf{X}$ 。

2) PLSI 和 NMF 优化同样的目标函数(K-L divergence), 但是 PLSI 有额外的正则化约束^[36]。两个模型的算法可以产生等价的解, 但两个算法本质上是不同的^[37]。

定理 3 NMF 和 PLSI 优化同样的目标函数。

定理 4 PLSI 的一个局部最优解是 NMF 以 K-L divergence 作为目标函数时的最优解, 反之亦然。即两个模型产生的最优解是等价的。

4 非负矩阵分解的应用

非负矩阵分解已被成功应用在数据分析的多个领域, 尤其是针对无监督学习。一般而言, NMF 的优点就是它能自动发现数据背后的隐含模式或者趋势。本部分简要回顾 NMF 在图像处理、数据聚类、半监督聚类和双聚类问题中的应用。

4.1 图像处理

正如之前提到的, NMF 在 Lee & Seung^[1] 的论文之后引发了广泛的注意。他们的论文主要讨论了 NMF 在图像处理上的应用。在图像处理问题中, 数据可以被表示成 $n \times m$ 的非负矩阵 \mathbf{X} , 每一列是 1 个图像, 由 n 个非负的像素值描述。NMF 可以将 \mathbf{X} 分解成 2 个因子矩阵 \mathbf{F} 和 \mathbf{G} , 使得 $\mathbf{X} \approx \mathbf{F}\mathbf{G}^T$ 。 \mathbf{F} 是基矩阵, 每一列可被认为是构成整体图像的 1 个部件, 比如对于人脸数据, 可以是鼻子、耳朵等。 \mathbf{G} 是编码矩阵, 每一行中的元素相当于赋给相应基向量的权重, 而 \mathbf{X} 中的图像根据不同的权重配比可以重建为基向量的线性组合。

总之, NMF 可以发现图像构成的隐含部件和图像通过这些部件重建的方式。事实上, 心理学上的研究也表明, 人的大脑就是将事物分解为若干部分来对事物进行认知的^[1]。但是进一步的研究表明标准的非负矩阵分解

模型并不一定得到恰当的“部分”(Part-of-whole)^[38-39],因此改进 NMF 模型,努力得到稀疏化的结果,以期得到更局部化的“部分”,成为 NMF 研究其中的一个重要方向。

4.2 数据聚类

NMF 最有趣和最成功的应用领域之一要属数据聚类了,应用领域包括文本、图像和生物数据。给定一个非负的 $n \times m$ 矩阵 \mathbf{X} ,每一列是 1 个数据样本,由 n 个特征描述,NMF 能够将其分解为 2 个因子矩阵 \mathbf{F} 和 \mathbf{G} 使得 $\mathbf{X} \approx \mathbf{F}\mathbf{G}^T$,其中 \mathbf{F} 是 $n \times r$ 的, \mathbf{G} 是 $m \times r$ 的, r 是预先设定的类数。 \mathbf{F} 的列可以被认为是聚类中心,而 \mathbf{G} 是聚类成员指示矩阵(Cluster membership indicator matrix)。换句话说,如果 G_{ik} 是行 $G_{i,:}$ 中最大的值,则数据样本 i 是属于聚类 k 的。

NMF 在数据聚类上的成功应用已经在多个领域得到验证,包括生物信息学^[25]、文本聚类^[21-23]和复杂网络上的社团结构探测等。

4.3 数据双聚类

双聚类(Bi-clustering, co-clustering)是由 Cheng & Church^[40]首次提出用于分析基因表达数据的。在基因表达数据中,许多基因在一些特定情况下会特异性表达,比如在肿瘤中或在组织生长的特定阶段。发现这种“基因-环境”的特定结构对于理解组织生长和代谢、肿瘤发生和诊断等问题具有重要价值,而传统聚类方法不适合发现此类结果。双聚类模型据此被提出,它可以同时对样本和特征 2 个维度进行聚类,从而发现此类局部结构。非负矩阵分解模型可应用于双聚类问题;比如对于肿瘤的基因表达数据矩阵 \mathbf{X} ,NMF 将其分解为 2 个因子矩阵 \mathbf{F} 和 \mathbf{G} ,肿瘤样本 i 是属于某类肿瘤 k 的,如果 G_{ik} 是行 $G_{i,:}$ 中最大的值,同时基因 j 是在该类肿瘤中特异表达的,如果 F_{jk} 是行 $F_{j,:}$ 中最大的值。这样就建立起了“基因-肿瘤”的子结构。但是这个结果是有不足之处的。可以看到,在此框架下,每个基因都会被分派到某类肿瘤中,即认为其在该类中特异表达。而实际上,引起肿瘤的基因在全部基因中所占的比重很小,大部分基因应该是不属于任何类别,即在任何肿瘤中都不特异表达。基于以上考虑,文献^[28-29]提出了二值矩阵分解模型(BMF):输入肿瘤样本的基因表达二值数据,BMF 将其分解为 2 个二值矩阵 \mathbf{F} 和 \mathbf{G} 。 \mathbf{F} 和 \mathbf{G} 可以明确地将特异性表达基因分配到某类肿瘤中,同时大部分基因所在的行 $F_{j,:}$ 为零向量,从而它们不属于任何类。下面的例子可以直观地说明该框架的有效性。

给定下面的二值目标矩阵 $\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ a & b & c & d & e & f & g & h \end{pmatrix}$, 可以看到矩阵中有 2 个子结构,一个在右上

角,另一个在左下角。BMF 模型给出如下的结果: $\mathbf{F} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}$; $\mathbf{G} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$ 。从而找到了子结构为

$\{1,2,3,4,f,g,h\}$ 和 $\{3,4,5,b,c,d\}$ 。重建的目标矩阵如下: $\mathbf{F}\mathbf{G}^T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$, 从中可以看到

2 个子结构更清晰地体现出来了。

4.4 以社团结构探测为例

复杂网络已经成为描述自然和社会中的系统结构的一个有力工具,是研究的热点。研究表明复杂网络经常

是具有社团结构的^[41-42]。复杂网络的一个社团结构是指网络顶点集的一个分划,每个子网络的内部连接是紧密的,而与外部的连接是疏松的。目前对社团结构问题的定义还没有统一的认识,在论文[43]中有一个关于社团结构问题的组合优化定义。许多证据表明,网络中的社团结构经常是有其特殊含义的。比如说,在生物网络中,社团结构经常是一个功能模块或者通路,而在社会网中,社团结构经常是由一群持相同观点的人组成,在科研合作网中,社团结构则经常是某一个科研领域等。文献中已经发展了大量有趣的模型来探测复杂社会网络中的社团结构^[44]。事实上,网络社团结构探测问题本质上是属于无监督学习的范畴,NMF 模型在其中有广阔的应用空间,比如下面这个为该问题设计的对称非负矩阵分解模型^[13]

$$\begin{aligned} \min \quad & \| \mathbf{X} - \mathbf{U}\mathbf{U}^T \|_F^2 \\ \text{s. t. } \quad & \mathbf{U} \geq 0, \sum_{j=1}^c \mathbf{U}_{ij} = 1, i = 1, 2, \dots, n \end{aligned} \quad (1)$$

其中 \mathbf{X} 是网络的邻接矩阵, \mathbf{U} 是节点的社团隶属度向量,即 \mathbf{U}_i 表示节点 i 对于社团 t 的隶属度(这里 $\sum_j \mathbf{U}_{ij} = 1$)。

但是,利用得到的隶属度向量并不能确定一个节点是否真的属于某个社团。也许隶属度 0.9 意味着相应的节点确实属于该社团,但是对于隶属度 0.5 就很难确定。明确节点是否真的属于某个社团显然是更有价值的^[45]。另外一个问题是隶属度向量不能将野点(Outerlier)和重叠点(Overlapping)区分开。比如说,一个节点的隶属度向量是 $\left[\frac{1}{c}, \frac{1}{c}, \dots, \frac{1}{c} \right]$,很难判断其含义是该节点以相同的概率属于各个社团,还是说该节点不属于任何一个社团。甚至后一种可能性更大。

一个解决此问题的方法是定义桥接性(Bridgeness)和中心度(Centrality)^[13]。如图 1 所示,基于桥接性和中心度的不同组合,可以将野点和重叠点区分开。具有高桥接性值和高中心度值的点更可能是沟通社团的桥点,而具有低中心度值和高桥接性值的点更可能是野点。点 i 的桥接性值可以

定义如下^[13] $G_1(i) = 1 - \sqrt{\frac{c}{c-1} \sum_{j=1}^c \left(\mathbf{U}_{ij} - \frac{1}{c} \right)^2}$ 。

注 图 1 度量了社团隶属度向量 \mathbf{U}_i ; 和参考向量 $\left[\frac{1}{c}, \frac{1}{c}, \dots, \frac{1}{c} \right]$ 之间的距离。有高的中心度,表明它们不能被清晰地分配到某个社团,但是它们是有影响力的节点,因而可认为它们是桥点。类似地,落在左下角的节点可能是社团中的边缘点或者是野点,而落在右下角的节点是社团的核心点。这种图示法是有局限性的,详见下文。

但是这种定义方法是有局限性的,有高 G_1 值的节点并不一定具有高的桥接性。例如,在一个四社团网络中,如果计算得到的点 i 的社团隶属度向量 $\mathbf{U}_{i..}$ 是 $\left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right]$,而点 j 的隶属度向量 $\mathbf{U}_{j..}$ 是 $\left[\frac{1}{2}, \frac{1}{2}, 0, 0 \right]$,相应的 $G_1(i)$ 和 $G_1(j)$ 就会分别是 1 和 0.42(甚至小于 0.5),这会给后续分析带来误导。而中心度的定义也很困难,就目前所知,如何定义节点在网络中的影响力还是一个未决的问题,仍是一个热点课题。不同的定义会导致非常不同的结果。最后,即便有好的定义,如何选取阈值来区分桥点和野点也是一个困难的问题。

基于以上考虑,笔者提出了对称二值矩阵分解模型(SBMF)来探测重叠社团结构^[18]。该模型可以明确地把节点分配到各个社团中,同时也能将野点和重叠点区分开。SBMF 可以表述如下:给定网络的邻接矩阵 \mathbf{A} ,找到二值矩阵 \mathbf{U} 使得 $\mathbf{A} \approx \mathbf{U}\mathbf{U}^T$ 。 \mathbf{U} 是节点的社团成员指示矩阵: $\mathbf{U}_i = 1$ 表示节点 i 是在社团 t 中的,0 表示不属于。如果节点 i 属于多个社团,则相应的第 i 行的和大于 1 ($\sum_j \mathbf{U}_{ij} > 1$)。反之,如果节点 i 是一个野点,则相应的第 i 行是零向量 ($\sum_j \mathbf{U}_{ij} = 0$)。同时假设只有很少的一部分节点是野点,从而只有很少的一部分向量是零向量 ($\sum_j \mathbf{U}_{ij} = 0$)。

综上所述,SBMF 模型可以被表述为如下非线性规划模型

$$\begin{aligned} \min \quad & \| \mathbf{A} - \mathbf{U}\mathbf{U}^T \|_1 + \sum_i \left(1 - \theta \left(\sum_j \mathbf{U}_{ij} \right) \right) \\ \text{s. t. } \quad & \mathbf{U}_{ij}^2 - \mathbf{U}_{ij} = 0, i = 1, 2, \dots, n, j = 1, 2, \dots, c \end{aligned} \quad (2)$$

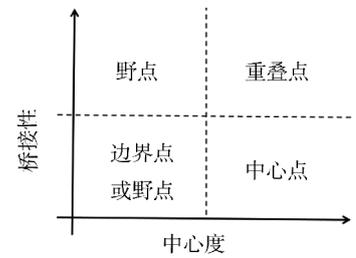


图 1 节点的中心度和桥接性
Fig. 1 Nodes' centrality versus nodes' bridgeness in complex networks

其中 Θ 是 Heaviside 步函数^[46], 对矩阵 \mathbf{X} , $\Theta(\mathbf{X})_{ij} := \begin{cases} 1, & \text{如果 } \mathbf{X}_{ij} > 0 \\ 0, & \text{如果 } \mathbf{X}_{ij} \leq 0 \end{cases}$ 。同时文献[18]还给出了一个选取恰当社团数目的指标。

NMF 模型还适用于半监督学习。尽管带约束的聚类问题已经在无监督学习领域中得到了广泛的研究。然而, 对于复杂网络中的社团结构探测问题, 如何将先验信息作为约束放到模型/算法中来引导探测过程仍然是一个有挑战性的问题。文献[17]提出一个半监督的网络社团结构探测框架。这个框架是根据先验信息来适当修改网络的邻接矩阵, 从而达到引导探测过程的目的。这种修改变可看做是对社团结构一致性矩阵的降噪和重现过程。简言之, 该框架兼顾了网络的拓扑结构性质和网络中点的功能信息, 从而能够大幅度提高社团结构探测模型/算法的成绩。将该框架应用于人工合成数据(GN 网络和 LFR 网络)和实际数据, 结果表明其确实能够显著提高探测成绩。文献[47]还提出了增强型的半监督学习框架, 能够进一步提高探测精度。

5 总结和展望

本文对非负矩阵分解的研究历史、模型、算法和应用进行了总结。简言之, 该模型相较于主成分分析具有更好的可解释性; 且模型简单, 易于理解操作, 可拓展性强, 通过选取不同的目标函数, 采用不同的约束可以实现不同的目的; 该模型和无监督学习领域中其它被广泛采用的模型关系紧密, 甚至将一些看似不相关的模型连接了起来; 它有广泛的应用空间, 数值表现优异。

最后给出几个值得进一步研究的方向作为本文的结尾。

1) 随着大数据时代的来临, 算法的高效稳健变得越来越重要, 有时候甚至可以牺牲一些精度来提高效率。因此如何设计更高效的算法来求解 NMF 是一个重要的问题^[48];

2) 数据分析越来越讲究任务分析, 所以如何将 NMF 和其它模型相结合也是一个有趣的问题;

3) 复杂网络的研究方兴未艾, 将研究视角从个体提升到了针对关系的研究, 如何更好地将复杂网络的自身特点嵌入 NMF 模型中亦是一个富有挑战性的问题。

参考文献:

- [1] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. *Natural*, 1999, 6755:788-791.
- [2] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[C]//Leen T K, Dietterich T G, Tresp V. In *Advances in Neural Information Processing Systems* 13. Cambridge, MA: MIT Press, 2000:556-562.
- [3] Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values[J]. *Environmetrics*, 1994, 5(2):111-126.
- [4] Jolliffe I T. *Principal Component Analysis*[M]. 2nd Edition. New York: Springer, 2002.
- [5] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction* [M]. 2nd Edition. New York: Springer, 2009.
- [6] Pearson K. On lines and planes of closest fit to systems of points in space[J]. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901, 2(11):559-572.
- [7] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. *Journal of educational psychology*, 1933, 24(6):417.
- [8] Hotelling H. Relations between two sets of variates[J]. *Biometrika*, 1936, 28(3/4):321-377.
- [9] Lawton W H, Sylvestre E A. Self modeling curve resolution[J]. *Technometrics*, 1971, 13(3):617-633.
- [10] Tropp J A. Literature survey: Non-negative matrix factorization [EB/OL]. (2003-03-24) [2013-07-25]. <http://users.cms.caltech.edu/~jtropp/notes/Tro03-Literature-Survey.pdf>.
- [11] Ho N D, Van Dooren P. Non-negative matrix factorization with fixed row and column sums[J]. *Linear Algebra and its Applications*, 2008, 429(5):1020-1025.
- [12] Vavasis S A. On the complexity of nonnegative matrix factorization[J]. *SIAM Journal on Optimization*, 2009, 20(3):1364-1377.
- [13] Nepusz T, Pécroci A, Négyessy L, et al. Fuzzy communities and the concept of bridgeness in complex networks[J]. *Physical Review E*, 2008, 77(1):016107.
- [14] 张忠元. 基于字典学习的网络社团结构探测算法[J]. *中国科学:信息科学*, 2011, 41(11):1343-1355.
Zhang Z Y. Community structure detection in social networks based on dictionary learning[J]. *Scientia Sinica Informationis*, 2011, 41(11):1343-1355.
- [15] Zhang Z Y. Community structure detection in social networks based on dictionary learning[J]. *Science China Information Sciences*, 2013, 56(7):1-12.
- [16] Psorakis I, Roberts S, Ebden M, et al. Overlapping com-

- munity detection using bayesian non-negative matrix factorization[J]. *Physical Review E*, 2011, 83(6):066114.
- [17] Zhang Z Y. Community structure detection in complex networks with partial back-ground information[J]. *Europhysics Letters*, 2013, 101(4):48005.
- [18] Zhang Z Y, Wang Y, Ahn Y Y. Overlapping community detection in complex net-works using symmetric binary matrix factorization[J]. *Physical Review E*, 2013, 87(6):062803.
- [19] Li S Z, Hou X W, Zhang H J, et al. Learning spatially localized, parts-based representation [C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001:207-212.
- [20] Cooper M, Foote J. Summarizing video using non-negative similarity matrix factorization[C]// Proc IEEE Workshop on Multimedia Signal Processing, 2002:25-28.
- [21] Pauca V P, Shahnaz F, Berry M W, et al. Text mining using non-negative matrix factorizations[C]//Proceedings of the Fourth SIAM International Conference on Data Mining, 2004.
- [22] Shahnaz F, Berry M W, Pauca V P, et al. Document clustering using nonnegative matrix factorization[J]. *Information Processing & Management*, 2006, 42(2):373-386.
- [23] Xu W, Liu X, Gong Y H. Document clustering based on non-negative matrix factorization [C]//SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. New York, USA; ACM Press, 2003:267-273.
- [24] Nielsen F A, Balslev D, Hansen L K. Mining the posterior cingulate: Segregation between memory and pain components[J]. *NeuroImage*, 2005, 27(3):520-532.
- [25] Brunet J P, Tamayo P, Golub T R, et L. Metagenes and molecular pattern discovery using matrix factorization[J]. *Proc Natl Acad Sci USA*, 2004, 101(12):4164-4169.
- [26] Pascual-Montano A, Carazo J M, Kochi K D, et al. Nonsmooth nonnegative matrix factorization (nsnmf)[J]. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(3):403-415.
- [27] Devarajan K. Nonnegative matrix factorization: An analytical and interpretive tool in computational biology [J]. *PLoS Comput Biol*, 2008, 4(7):e1000029.
- [28] Zhang Z Y, Li T, Ding C, et al. Binary matrix factorization for analyzing gene expression data [J]. *Data Min Knowl Discov*, 2010, 20(1):28-52.
- [29] Zhang Z Y, Li T, Ding C, et al. Binary matrix factorization with applications [C]//ICDM '07 Proceedings of the 2007 Seventh IEEE International Conference on Data Mining. Washington, DC, USA; IEEE Computer Society, 2007:391-400.
- [30] Zhang Z Y. Nonnegative matrix factorization: models, algorithms and applications[J]. *Data Mining: Found & Intell Paradigms, ISRL*, 2012, 24:99-134.
- [31] Lin C J. Projected gradient methods for nonnegative matrix factorization[J]. *Neural computation*, 2007, 19(10):2756-2779.
- [32] Mairal J, Bach F, Ponce J, et al. Online learning for matrix factorization and sparse coding [J]. *The Journal of Machine Learning Research*, 2010, 11:19-60.
- [33] Zhang Z Y. Divergence functions of non negative matrix factorization: a comparison study [J]. *Communications in Statistics-Simulation and Computation*, 2011, 40(10):1594-1612.
- [34] Gonzalez E F, Zhang Y. Accelerating the Lee-Seung algorithm for non-negative matrix factorization[R]. Houston, TX; Department of Computational and Applied Mathematics, Rice University, 2005.
- [35] Ding C, He X F, Simon H D. On the equivalence of non-negative matrix factorization and spectral clustering [C]// In SIAM International Conference on Data Mining, 2005:606-610.
- [36] Ding C, Li T, Peng W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing [J]. *Comput Stat Data Anal*, 2008, 52(8):3913-3927.
- [37] Zhang Z Y, Ding C, Tang J. Note on algorithm differences between nonnegative matrix factorization and probabilistic latent semantic indexing [J]. *Journal of Convergence Information Technology*, 2011, 6(9):210-219.
- [38] Li S Z, Hou X W, Zhang H J, et al. Learning spatially localized, parts-based representation [C]//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Kauai, Hawaii, 2001:207-212.
- [39] Hoyer P O, Dayan P. Non-negative matrix factorization with sparseness constraints [J]. *Journal of Machine Learning Research*, 2004, 5:1457-1469.
- [40] Cheng Y Z, Church G M. Biclustering of expression data [C]//Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, 2000:93-103.
- [41] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. *Nature*, 2005, 435:814-818.
- [42] Girvan M, Newman M E J. Community structure in social and biological networks [J]. *Proceedings of the National Academy of Sciences*, 2002, 99(12):7821-7826.
- [43] Zhang X S, Li Z, Wang R S, et al. A combinatorial model and algorithm for globally searching community structure in complex networks [J]. *Journal of Combinatorial Opti-*

- mization, 2012, 23(4):425-442.
- [44] Fortunato S. Community detection in graphs[J]. Physics Reports, 2010, 486(3):75-174.
- [45] Ahn Y, Bagrow J, Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature, 2010, 466:761-764.
- [46] Abramowitz M, Stegun I A. Handbook of mathematical functions with formulas, graphs and mathematical tables [S]//Abramowitz M, Stegun I A. National Bureau of Standards Applied Mathematics Series 55. tenth printing. Washington, USA: United States Department of Commerce, 1972.
- [47] Zhang Z Y, Sun K D, Wang S Q. Enhanced community structure detection in social networks[EB/OL]. (2013-05-14)[2013-07-25]. <http://arxiv.org/pdf/1210.2473.pdf>.
- [48] Liu C, Yang H C, Fan J L, et al. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce[EB/OL]. (2010-02-10) [2013-07-25]. <http://research.microsoft.com/pubs/119077/dnmf.pdf>.

Operations Research and Cybernetics

Nonnegative Matrix Factorization: Model, Algorithms and Applications

ZHANG Xiang-sun¹, ZHANG Zhong-yuan²

(1. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190;

2. School of Statistics and Mathematics, Central University of Finance and Economic, Beijing 100081, China)

Abstract: Nonnegative Matrix Factorization (NMF) is becoming one of the most popular models in data mining society recently. NMF can extract hidden patterns from a series of high-dimensional vectors automatically, and has been applied for dimensional reduction, unsupervised learning (image processing, clustering and co-clustering, etc.) and prediction successfully. This paper surveys NMF in terms of the research history, model formulation, algorithms and applications. In summary, NMF has good interpretability, is very flexible, has a close relationship with the existing state of the art unsupervised learning models and a variety of applications. In addition, as a developing technology, there are still many interesting open issues remained unsolved and waiting for research from different perspectives.

Key words: nonnegative matrix factorization; multiplicative update algorithms; K-means; PLSI; image processing; clustering; community structure detection

(责任编辑 黄颖)