ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 6,Nov. 2017,pp. 1216-1222
DOI:10. 16337/j. 1004-9037. 2017. 06. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:+86-025-84892742

基于稀疏约束非负矩阵分解的 K-Means 聚类算法

韩素青　贾　茹

(太原师范学院计算机科学与技术系,太原,030619)

摘　要: 为了提高 K-Means 聚类算法在高维数据下的聚类效果,提出一种基于稀疏约束非负矩阵分解的 K-Means 聚类算法。该算法在最优保持原始数据本质的前提下,通过在非负矩阵分解过程中对基矩阵列向量施加 l_1 与 l_2 范数稀疏约束,首先挖掘嵌入在高维数据中的低维数据结构,实现高维数据的低维表示,然后利用在低维数据聚类中性能良好的 K-Means 算法对稀疏降维后的数据进行聚类。实验结果表明提出的算法可行,并且在处理高维数据上有效。

关键词: 高维数据;非负矩阵分解;稀疏约束;k-means 聚类

中图分类号: TP274　　　**文献标志码:** A

K-Means Clustering Algorithm Based on Non-negative Matrix Factorization with Sparseness Constraints

Han Suqing, Jia Ru

(Department of Computer Science and Technology, Taiyuan Normal University, Taiyuan, 030619, China)

Abstract: To improve the quality of K-Means clustering in high-dimensional data, a K-Means clustering algorithm is presented based on non-negative matrix factorization with sparseness constraints. The algorithm finds the low dimensional data structure embedded in high-dimensional data by adding l_1 and l_2 norm sparseness constraints to the non-negative matrix factorization, and achieves low dimensional representation of high dimensional data. Then the K-Means algorithm, which is the high performance clustering algorithm in low dimensional data, is used to cluster the low dimensional representation of high dimensional data. The experimental results show that the proposed algorithm is feasible and effective in dealing with high-dimensional data.

Key words: high dimensional data; non-negative matrix factorization; sparse constraint; K-Means clustering

引　言

聚类分析是数据挖掘领域中一个热门的研究方向,其主要思想是将数据对象划分成不同的簇,使得同一簇中的数据对象具有较高的相似度,而不同簇中的数据对象具有较低的相似度。目前基于划分、层次、密度、网格和神经网络等的诸多聚类算法在商务、市场分析、生物学以及文档分析等领域有广泛的应

基金项目: 国家自然科学基金(61273294)资助项目。
收稿日期: 2017-04-01;**修订日期:** 2017-06-19

用。这些聚类算法在处理小规模数据和低维数据时一般都有比较好的效果，但是随着数据量和数据维数的逐渐升高，性能会急剧下降，因此不适用于对高维数据进行聚类分析[1]。

K-Means[2-3]是基于划分的聚类算法，是一种有着广泛应用，且具有代表性的聚类方法，具有简单、易于实现、可伸缩和高效等优点。但是，随着数据维数的逐渐增高，可能会受到“维数灾难”的影响，难以获得好的聚类结果。这意味着，K-Means 聚类算法在处理数据维数不断升高的聚类问题时，对数据进行有效降维是必要的。

矩阵分解[4]对数据降维非常有效，通过矩阵分解得到原数据矩阵的低秩逼近，是对原始数据内在结构特征的刻画。目前，利用矩阵分解来解决实际问题的数据分析方法很多，如主成分分析（Principal component analysis，PCA）、独立成分分析（Independent component analysis，ICA）和奇异值分解（Singular value decomposition，SVD）等。这些方法的共同特点是，分解后矩阵因子中的元素可正可负。从数学的角度看，分解结果中存在负值是正确的，但负元素在实际问题中往往是没有意义的。

非负矩阵分解[5-6]（Non-negative matrix factor-orization，NMF）作为矩阵分解的一种特殊方法，由于其可以在保持原始数据本质的前提下，实现高维数据的低维表示，在图像分析、文本聚类、数据挖掘、语音处理、机器人控制、生物医学工程和化学工程等领域有着广泛的应用。

NMF 最早由 Lee D 和 Seung H S 提出，发表于《Nature》，是一种在矩阵中所有元素均为非负数约束下的矩阵分解。它具有简洁明了，易于操作，分解结果可解释、占用存储空间少，并且可以从高维数据中发现隐藏其中的有意义的低维结构[7]等优点。之后，Ding C，He Xiaofeng 和 Simon H D[8]系统分析和研究了 NMF 的对称扩展形式

$$\boldsymbol{V}=\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$$

和加权扩展形式

$$\boldsymbol{V}=\boldsymbol{H}\boldsymbol{S}\boldsymbol{H}^{\mathrm{T}}$$

并证明了对称扩展形式等价于核 K-Means 聚类。

NMF 分解后的基矩阵和系数矩阵虽然本身具有较好的稀疏[9-10]性，但仍然需要耗费巨大的存储空间，因此，许多学者在 NMF 基础上，对数据的稀疏性做了进一步的研究。Hoyer[11]首次提出 NMF 与稀疏编码结合构造非负稀疏编码（Non-negative matrix factorization with sparseness constraints，NMFSC）算法的思想，其在稀疏约束条件下，通过 NMF 获得稀疏的基矩阵，减少了存储空间。Li 等[12]在标准 NMF 基础上，通过对基矩阵增加空间局部化限制，构造了 LNMF 算法。Xu 等[13]提出了一种基于信息检索的受限非负矩阵分解。黄钢石等[14]提出了一种获取潜在语义的受限非负矩阵分解方法（Constrained factorization method for non-negative matrix，CNMF）算法。胡学考等[15]利用少量标注样本和大量未标注样本，构造了施加稀疏约束的非负矩阵分解算法。刘建伟等[16]总结了各种正则化稀疏模型，分析了各种稀疏模型被提出的原因、所具有的优点、适宜解决的问题及其模型的具体形式等。虽然上述方法都比较好地解决了非负矩阵分解的稀疏性增强问题，但大多研究都集中在对 NMF 分解后的矩阵进行稀疏约束。

本文基于 NMF 分解，研究 K-Means 聚类算法在高维数据聚类中的应用。提出了基于稀疏约束非负矩阵分解的 K-Means 聚类算法（Non-matrix factorization with sparse constraint for K-means clustering，NMFS-K）。算法对原始高维数据构成的矩阵 $\boldsymbol{V}$ 首先进行 NMF，即，近似计算 $\boldsymbol{V}\approx\boldsymbol{W}\boldsymbol{H}$，其中，$\boldsymbol{W}$ 为基矩阵，$\boldsymbol{H}$ 为权矩阵。基于这种分解，$\boldsymbol{V}$ 的列向量可以解释为对 $\boldsymbol{W}$ 中所有列向量的加权和。为了减少存储空间，同时获得好的聚类结果，在非负矩阵分解过程中，对基矩阵 $\boldsymbol{W}$ 的每一列 $\boldsymbol{w}_i$ 施以 l_1 与 l_2 范数的稀疏约束，并且为了获得稳定的基矩阵 $\boldsymbol{W}$，对所得的 $\boldsymbol{W}$ 的列进行了归一化处理；最后利用 K-Means 聚类算法，对稀疏降维后的低维数据进行聚类。

1 基本概念

1.1 K-Means 算法

K-Means 算法由 MacQueen 于 1967 年首次提出，是一种常用的基于划分的聚类分析方法，其核心思想是找出 k 个聚类中心，则

$$\Omega=\{\omega_1,\omega_2,\cdots,\omega_k\}$$

并使每一个数据点 $\boldsymbol{x}$ 和与其最近的聚类中心 ω_i 的距离平方和最小。

K-Means 算法的主要优点是简单、快速和有效，但仍存在一些不足：(1)聚类结果易受 k 个初始中心选取的影响；(2)需要根据初始聚类中心确定初始划分，并通过对初始划分进行优化求解，容易陷入局部最优；(3)计算数据之间的距离时需要遍历中心点，当数据规模较大时，算法的时间开销比较大。

1.2 非负矩阵分解

NMF 将高阶原始矩阵分解为两个阶数较低的非负矩阵乘积的一般形式[17-18]，即

$$\boldsymbol{V}=\boldsymbol{WH}+\boldsymbol{E}\Rightarrow\boldsymbol{V}\approx\boldsymbol{WH}$$

式中：$\boldsymbol{V}\in\mathbf{R}^{m\times n}$ 为待分解矩阵，$\boldsymbol{W}\in\mathbf{R}^{m\times r}$ 和 $\boldsymbol{H}\in\mathbf{R}^{r\times n}$ 为分解后的两个非负矩阵，$\boldsymbol{E}$ 为逼近误差，要求$\boldsymbol{E}$尽可能小并且快速收敛。r 的选取根据实际情况设置，一般要求 $r\ll\min\{m,n\}$。

NMF 是 NP-hard 问题，可转化为优化问题求解，在分解过程中通过不断迭代优化的方法交替求解 $\boldsymbol{W}$ 和 $\boldsymbol{H}$。

常用的两种目标函数为

$$\min_{\boldsymbol{WH}}=\boldsymbol{V}-\boldsymbol{WH}=\sum_{i,j}(V_{ij}-(WH)_{ij})^2$$

和

$$\min_{\boldsymbol{W},\boldsymbol{H}}\sum_{i,j}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}}-V_{ij}+(WH)_{ij}\right)$$

分别基于欧式距离和 K-L 散度定义。

由于$\boldsymbol{W}$,$\boldsymbol{H}$ 非凸，所以对应的最优解通过迭代公式进行求解，当$\boldsymbol{V}$-$\boldsymbol{WH}$误差小于某一个数或者达到最大迭代次数时，算法停止，获得最优解 $\boldsymbol{W}$ 和 $\boldsymbol{H}$。

1.3 向量的稀疏度

稀疏约束 NMF 指在迭代过程中对基矩阵 $\boldsymbol{W}$ 的列向量或权矩阵 $\boldsymbol{H}$ 的行向量进行稀疏约束的非负矩阵分解。由于向量的稀疏性与向量中非零分量所占的比例有关，而与分量的符号和顺序无关，因此，向量的稀疏性可以用稀疏度来度量。Hoyer 给出的稀疏度[11]为

$$sp(\boldsymbol{x})=\frac{\sqrt{m}-\dfrac{\boldsymbol{x}_1}{\boldsymbol{x}_2}}{\sqrt{m}-1}\tag{1}$$

式中：m 为向量 $\boldsymbol{x}$ 的维数，$sp(\boldsymbol{x})\in[0,1]$，$L_1=\boldsymbol{x}_1=\sum_{i=1}^{n}|x_i|$，$L_2=\boldsymbol{x}_2=[\sum_{i=1}^{n}|x_i|^2]^{1/2}$。其中：$l_1$ 范数$\boldsymbol{x}_1$ 用来控制稀疏解的产生，通过最小化 l_1 范数$\boldsymbol{x}_1$，可以找到最少且最优的稀疏特征项。l_2 范数$\boldsymbol{x}_2$ 用来约束模型参数，可以抑制过拟合，增强模型的泛化能力。

从式(1)可以看出，稀疏度 $sp(\boldsymbol{x})$ 的值越大，向量中分量的分布越零散，分量值间的差异越大；稀疏度 $sp(\boldsymbol{x})$ 的值越小，向量中的分量分布越均匀，分量值在某一固定值上下波动的幅度就越小。因此，用式(1)对向量的稀疏性进行刻画是一种合理的选择。

1.4 稀疏约束 NMF 算法

将 NMF 应用于高维数据聚类时,因为基矩阵 $\boldsymbol{W}$ 反映了原始高维数据的主要特征,因此将基矩阵 $\boldsymbol{W}$ 解释为高维数据的原型矩阵。对 $\boldsymbol{W}$ 施加稀疏约束(即,对 $\boldsymbol{W}$ 的每一列 $\boldsymbol{w}_i$ 施加稀疏约束),可以在最优保持原始信息的条件下,减少基矩阵 $\boldsymbol{W}$ 的存储空间,为 K-Means 聚类奠定基础。

稀疏约束 NMF 算法通过迭代求解最优 $\boldsymbol{W}$ 和 $\boldsymbol{H}$。分解过程采用梯度下降策略,利用乘性和加性迭代更新规则来获得基矩阵 $\boldsymbol{W}$ 和权矩阵 $\boldsymbol{H}$。

这里,加性迭代规则指

$$\boldsymbol{W}=\boldsymbol{W}-\mu_w(\boldsymbol{WH}-\boldsymbol{V})\boldsymbol{H}^{\mathrm{T}} \tag{2}$$

乘性迭代规则指

$$\begin{aligned}\boldsymbol{W}&\leftarrow\boldsymbol{W}.*(\boldsymbol{VH}^{\mathrm{T}})./(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{HH}^{\mathrm{T}})\\ \boldsymbol{H}&\leftarrow\boldsymbol{H}.*(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V})./(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{WH})\end{aligned} \tag{3}$$

式中:μ_w 为迭代步长;“. *”和“./”分别代表两个矩阵元素的点乘与点除。

加性迭代和乘性迭代的作用在于,加性迭代保证只要所有元素被设定为足够小的正数,那么欧氏距离就会随着加性迭代而减小;而乘性迭代规则保证迭代的每一步的结果均为正数,即可以保证迭代的每一步获得的矩阵是非负的。

在对基矩阵 $\boldsymbol{W}$ 进行稀疏约束的过程中,有可能引起 NMF 识变率降低的问题,因此,需要对基矩阵 $\boldsymbol{W}$ 进行归一化处理以获得稳定结果,进而提高 NMF 的识变率,并且在保持一定识别率的基础下使基矩阵 $\boldsymbol{W}$ 尽可能的稀疏。

基矩阵 $\boldsymbol{W}$ 的归一化描述为

$$\boldsymbol{W}=\left(\frac{\boldsymbol{w}_1}{\boldsymbol{W}_2},\frac{\boldsymbol{w}_2}{\boldsymbol{W}_2},\cdots,\frac{\boldsymbol{w}_r}{\boldsymbol{W}_2}\right) \tag{4}$$

其本质是对基矩阵 $\boldsymbol{W}$ 的每一列 $\boldsymbol{w}_i$ 进行 2-范数归一化。

稀疏约束非负矩阵分解算法(Non-negative matrix factorization with sparseness constraints, SNMF[10]),描述如下。

输入:原始矩阵 $\boldsymbol{V}$,预设降维后的维数 r,稀疏度 $sp(\boldsymbol{x})$,迭代步长 μ_w。

输出:满足稀疏度 $sp(\boldsymbol{x})$ 的基矩阵 $\boldsymbol{W}$。

(1) 初始化 $\boldsymbol{W}$ 和 $\boldsymbol{H}$ 为两个随机的非负矩阵。

(2) 对 $\boldsymbol{W}$ 的每一列 $\boldsymbol{w}_i$ 进行稀疏约束,然后将 $\boldsymbol{w}_i$ 投影为非负值,使 $\boldsymbol{w}_i$ 的 l_2 范数不变,而 l_1 的范数根据预置的稀疏度 $sp(\boldsymbol{x})$ 由式(1)计算。

(3) 令

$$\boldsymbol{W}=\boldsymbol{W}-\mu_w(\boldsymbol{WH}-\boldsymbol{V})\boldsymbol{H}^{\mathrm{T}}$$

并将 $\boldsymbol{w}_i$ 投影为非负值,使 $\boldsymbol{w}_i$ 的 l_2 范数不变,而 l_1 的范数根据预置的稀疏度 $sp(\boldsymbol{x})$ 由式(1)计算。

(4) 令

$$\begin{aligned}\boldsymbol{W}&\leftarrow\boldsymbol{W}.*(\boldsymbol{VH}^{\mathrm{T}})./(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{HH}^{\mathrm{T}})\\ \boldsymbol{H}&\leftarrow\boldsymbol{H}.*(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V})./(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{WH})\end{aligned}$$

并将 $\boldsymbol{w}_i$ 投影为非负值,使 $\boldsymbol{w}_i$ 的 l_2 范数不变,l_1 的范数根据预置的稀疏度 $sp(\boldsymbol{x})$ 由式(1)计算。

(5) 若 $d=\boldsymbol{V}-\boldsymbol{WH}^2<\tau$,或迭代次数超过限制,则终止迭代,否则跳至步骤(2)直至迭代结束。

(6) 利用式(4)对 $\boldsymbol{W}$ 的列进行归一化,获得基矩阵 $\boldsymbol{W}$。

2 基于稀疏约束 NMF 的 K-Means 聚类算法

本节基于 SNMF 算法和 K-Means 聚类算法设计基于稀疏约束非负矩阵分解的 K-Means 聚类算

法，简称 NMFS-K 算法。

NMFS-K 算法

输入：数据集 X，预设的聚类数 k。

输出：聚类结果 C。

(1) 将数据集 X 用矩阵 $\boldsymbol{V}_{m\times n}$ 表示，其中，每一列为一个数据样本，由 m 个特征描述，共 n 个数据样本。

(2) 调用 SNMF 算法得到稀疏约束 NMF 后的基矩阵 $\boldsymbol{W}$。

(3) 将得到的 $\boldsymbol{W}$ 每一列 $\boldsymbol{w}_i$，$i=1,2,\cdots,k$ 作为一个初始聚类中心。

(4) 迭代

若 $h_{ij}(i=1,2,\cdots,k,j=1,2,\cdots,n)$ 是列 $\boldsymbol{H}_j$ 中的最大值，则将第 j 个数据样本归到聚类中心为 $\boldsymbol{w}_i$ 的簇中，计算每个簇的簇内平均值，将其作为该簇的新的类中心，直到所有类中心不再改变，或达到最大迭代次数，算法停止。

(6) 返回聚类结果 $C^*=(C_1,C_2,\cdots,C_k)$。

3 实验与结果分析

3.1 实验数据与实验环境

本文从 UCI 数据集中选取 Iris，Wine，Vote，German，Wpbc，Landsat，Spambase，Sonar，Dna 和 Msplice 等 10 个经典的数据集进行测试，数据集分布如表 1 所示。

表 1 UCI 数据集

Tab. 1 UCI data set

Name	Attributes	Instances	Class
Iris	4	150	3
Wine	13	178	3
Aistralian	15	690	2
Vote	16	435	2
German	24	1000	2
Wpbc	33	198	2
Spambase	57	4061	2
Sonar	60	208	2
Dna	180	2000	3
Msplice	240	3175	3

实验在 Intel 联 i7-4790CPU，4 GB 存储，Windows 7 操作系统，Matlab 2013a MyEclipse 环境下进行。

3.2 实验结果分析

基于表 1 所示的 UCI 数据集，对 K-Means 算法与 NMFS-K 算法的聚类结果进行对比分析，分别用分类准确率[19]与时间、空间复杂度对聚类结果进行评价。这里，分类准确率指

$$\mathrm{CA}=\frac{\sum_{i=1}^{k}a_i}{n}$$

式中：k 为聚类的个数，a_i 为正确聚类时对应类别 C_i 中样本的个数，n 为样本总数。显然，CA 的值越大，

表明聚类结果越好。

关于分类准确率。在表 1 所示的数据集上分别独立运行 K-Means 算法与 NMFS-K 算法各 10 次,并设定迭代次数不超过 100 次,然后取平均值进行比较。为避免算法陷入局部最优或无限迭代,基矩阵的稀疏度设为 0.3;同时,为避免耗费时间过长,迭代步长 μ_w 设为 0.001。实验结果如表 2 所示。

表 2 分类准确率

Tab. 2 Classification accuracy

Name	K-Means	NMFS-K
Iris	0.893	0.866
Wine	0.567	0.642
Aistralian	0.559	0.630
Vote	0.866	0.890
German	0.675	0.726
Wpbc	0.601	0.651
Spambas	0.635	0.780
Sonar	0.552	0.572
Dna	0.752	0.603
Msplice	0.611	0.673

从表 2 可以看出,对于小的、低维的数据集,NMFS-K 算法的优势不足,但是,随着数据维数的增加,相对于 K-Means 算法,NMFS-K 算法整体上有一定的优势。时间复杂度方面,NMFS-K 算法的时间复杂度为 $O(\mathrm{nrt}(s+k))$(s 为稀疏约束所用的时间),K-Means 算法的时间复杂度为 $O(nmkt)$。原因是 NMFS-K 算法需要进行 NMF 降维和稀疏约束,所以相比于 K-Means 算法,NMFS-K 算法的运行时间略长。

空间复杂度方面,NMFS-K 算法的空间复杂度为 $O(0.3(2k)*n)$,其中,0.3 是稀疏约束带来的空间效率,k 为原数据 NMF 降维后的维数,同时也是 NMFS-K 算法聚类的个数;K-Means 算法的空间复杂度为 $O((m+k)*n)$。显然 NMF 降维和稀疏约束发挥了作用,所以相比于 K-Means 算法,NMFS-K 算法具有明显的空间效率。

4 结束语

网络时代,许多数据分析方法都需要借助矩阵的形式来表达和处理,NMF 思想为研究者提供了高维数据分析处理的新思路。本文针对 K-Means 聚类算法随数据维数增高聚类效果降低的问题,提出了 NMFS-K 算法。该算法采用稀疏约束非负矩阵分解对高维数据进行维数约简,并利用低维性能良好的 K-Means 聚类算法对降维后的数据进行聚类,获得了比较好的结果,为高维数据聚类分析提供了一种可行且有效的解决方案,拓展了 K-Means 聚类算法的应用空间。

参考文献:

[1] 梁吉业,冯晨娇,宋鹏. 大数据相关分析综述[J]. 计算机学报,2016,39 (1):1-18.
Liang Jiye, Feng Chenjiao, Song Peng. A survey on correlation analysis of big data[J]. Chinese Journal of Computers,2016,39(1):1-18.

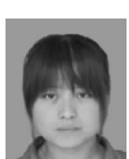
[2] 黄韬,刘胜辉,谭艳娜. 基于 k-means 聚类算法的研究[J]. 计算机技术与发展,2011(7):54-57,62.
Huang Tao,Liu Shenghui,Tan Yanna. Research of clustering algorithm based on K- means[J]. Computer Technology and development,2011(7):54-57,62.

[3] 周爱武,于亚飞. K-Means 聚类算法的研究[J]. 计算机技术与发展,2011(2):62-65.
Zhou Aiwu,Yu Yafei. The research about clustering algorithm of k-means[J]. Computer Technology and Development,2011

(2):62-65.
[4] 徐成贤,徐宗本.矩阵分析[M].西安:西北工业大学出版社,1991.
Xu Chengxian, Xu Zongben. Matrix Analysis[M]. Xi'an: North Western Poly Technical University Press, 1991.
[5] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788-791.
[6] 李乐.章毓晋.非负矩阵分解算法综述[J].电子学报,2008,36(4):737- 713.
Li Le, Zhang Yujin. A survey on algorithms of non-negative matrix f-actorization[J]. Acta Electronica Sinica, 2008, 36(4): 737-743.
[7] Zhou G, Cichocki A, Xie S. Fast non-negativ-e matrix/ tensor fac- torization based on low-rank approxi-mation[J]. IEEE Trans Signal Processing, 2012, 60(6): 292-294.
[8] Ding C, He X, Simon H D. On the equivalence of nonnegative matrix factorization and spectral clustering[C]//Proceedings of the 2005 SIAM International Conference on Data Mining. [S. l.]: Society for Industrial and Applied Mathematics, 2005: 606-610.
[9] 张荣,孙权森. 基于核稀疏保持投影的典型相关分析算法[J]. 数据采集与处理,2017,32(1):111-118.
Zhang Rong, Sun Quansen. Canonical correlation analysis algorithm based on kernal sparsity preserve projection[J]. Journal of Data Acquisition and Processing, 2017, 32(1): 111-118.
[10] 刘站琳,林小竹.基于稀疏约束的非负矩阵分解算法研究[J].北京印刷学院学报,2015(4):53-58.
Liu Zhanlin, Lin Xiaozhu. Sparse constraint non-negative matrix factorization[J]. Journalof Beijing Institute of Graphic Communication, 2015(4): 53-58.
[11] Hoyer. Non-negative matrix factorization with sparseness constrains[J]. Journal of Machine Learning Research, 2004(5): 1457-1469.
[12] Li Stan, Hou Xinwen, Zhang Hangjiang, et al. Learning spatially loca-lized, partsed representation[C]//Pro Cof Comp Visionand Pattern Recog. Los Alamitos, California, USA: [s. n.], 2001: 207-212.
[13] Xu Baowen, Lu Jianjiang, Huang Gangshi. A constrainednon-negatitive matrix factorization in information retrievval[C]//Proc of the 2003IEEE Int Confon Information Reuse and Integration. Nevada, USA: IEEE, 2003: 273-277.
[14] 黄钢石,张亚飞,陆建江,等.一种受限非负矩阵分解法[J].东南大学学报(自然科学版),2004,34(2):189-193.
Huang Gangshi, Zhang Yafei, Lu Jianjiang, et al. Constrained factoriza-tion method for non-negative matrix[J]. Journal of Southeast University (Natural Science Edition), 2004, 34(2): 189-193.
[15] 胡学考,孙福明,李豪杰.基于稀疏约束的半监督非负矩阵分解算法[J].计算机科学,2015,42(7):280-284,304.
Hu Xuekao, Sun Fuming, Li Haojie. Constrained non-negative matri-X factorization with sparseness for image representation [J]. Computer Science, 2015, 42(7): 280-284, 304.
[16] 刘建伟,崔立鹏,刘泽宇,等. 正则化稀疏模型[J].计算机学报,2015(7):1307-1325.
Liu Jianwei, Cui Lipeng, Liu Zeyu, et al. Survey on the regularized sparse models[J]. Computer Science, 2015(7): 1307-1325.
[17] 张雄伟,李轶南,时文华,等. 非负组合模型及其在声源分离中的应用[J]. 数据采集与处理,2017,22(2):266-277.
Zhong Xiongwei, Li Yinan, Shi Wenhua, et al. Non-negative compositional models and its application in acoustic source separation [J]. Journal of Data Acquisition and Processing, 2017, 22(2): 266-277.
[18] 张永亮.矩阵的非负分解算法及应用[D].杭州:浙江大学,2008.
Zang Yongliang. Non-negative matrix algorithm with application[D]. Hangzhou: Zhengjiang University, 2008.
[19] Liang Jiye, Bai Liangdang, Chuang Yin, et al. The K-means-type algorithms versusim balanced data distributions[J]. IEEE Transition Fuzzy Systems, 2012, 20(4): 728-745.

作者简介:

韩素青(1964-),女,博士,研究方向:数据挖掘、机器学习、粗糙集理论,E-mail:hansuqing@tynu.edu.cn。



贾　茹(1991-),女,硕士,研究方向:智能计算与数据建模,E-mail:18435172732@163.com。