第 34 卷 第 16 期 计 算 机 工 程 2008 年 8 月
Vol.34 No.16 Computer Engineering August 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)16—0191—03 文献标识码：A 中图分类号：TP391

NMF 初始化研究及其在文本分类中的应用

翟亚利，吴 翊
(国防科学技术大学理学院，长沙 410073)

摘 要：对非负矩阵分解的初始化进行研究，提出针对文本分类的主成分分析(PCA)、有监督 PCA(SPCA)和模糊 C 平均 3 种初始化方法并进行了实验。多类文本分类的实验结果表明，这些方法有效地解决了初值对结果的影响问题，不同程度地提高了文本分类结果，其中 SPCA 优于其他 2 种方法。

关键词：非负矩阵分解；模糊 C 平均；文本分类

Study of Non-negative Matrix Factorization Initialization and Its Application to Text Classification

ZHAI Ya-li, WU Yi
(College of Sciences, National University of Defence Technology, Changsha 410073)

【Abstract】 The initialization of Non-negative Matrix Factorization(NMF) has studied in this paper. There are three methods of initialization PCA, supervised PCA(SPCA) and Fuzzy C-Mean(FCM) are reported for text classification. Experimentsal results of multi-class text classification indicate that the three methods effectively solve the problem of results effected by initialized values, and improve the text classification results. The SPCA of the three methods is best.

【Key words】 Non-negative Matrix Factorization(NMF); Fuzzy C-Mean(FCM); text classification

1 概述

1999 年 D.D.Lee 和 H.S.Seung 在著名的科学杂志《Nature》上提出了一种新的矩阵分解思想——非负矩阵分解(Non-negative Matrix Factorization, NMF)及其算法[1]，即 NMF是在矩阵中所有元素均为非负数约束条件之下的矩阵分解方法。该思想为人类处理大规模数据提供了一种新的途径，具有实现上的简便性、分解形式和分解结果上的可解释性，以及占用存储空间少等诸多优点，体现了人类思维中“局部构成整体”的观念。除了目前图像分析、文本聚类(数据挖掘)、语音处理等的应用领域外，该方法还将可以用于维数消减、数据压缩等方面。

NMF、局部非负矩阵分解(LNMF)等各种类似方法在信息检索领域有广泛的应用[2-3]，用于获取文本集中的潜在语义。本文研究NMF迭代算法中的初始化问题，针对文本分类使用 3 种初始化方法——主成分分析(PCA)、有监督PCA (supervised PCA, SPCA)和模糊C平均(Fuzzy C-Mean, FCM)，使用向量空间模型进行文本分类实验，在测试集按时间选择的多类文本分类中，对词条文本矩阵使用NMF进行降维，实验结果表明：这 3 种初始化方法有效地解决了分类结果受初始值影响的问题，与随机初始化相对比都不同程度地提高了文本分类精度，其中，SPCA优于其他 2 种方法。

2 非负矩阵分解

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 $\boldsymbol{X}_{n\times m}$ 表示。该数据矩阵中各个元素都是非负的，表示为 $\boldsymbol{X}\geq 0$。对矩阵 $\boldsymbol{X}_{n\times m}$ 进行线性分解，有

$$\boldsymbol{X}_{n\times m}\approx \boldsymbol{W}_{n\times r}\times \boldsymbol{H}_{r\times m} \tag{1}$$

其中，$\boldsymbol{W}_{n\times r}$ 称为基矩阵；$\boldsymbol{H}_{r\times m}$ 称为系数矩阵，要求 $\boldsymbol{W}\geq 0$，$\boldsymbol{H}\geq 0$，r 满足 $(n+m)r<nm$。非负矩阵分解是一个 NP 问题，通常采用迭代算法寻求在非负性约束下使得目标函数最优的 $\boldsymbol{W}$ 和 $\boldsymbol{H}$。

采用概率模型和加性噪声模型[4]可以得到对非负矩阵分解的近似度的度量函数，即损失函数。

概率模型为

$$\boldsymbol{X}_{n\times m}=\boldsymbol{W}_{n\times r}\boldsymbol{H}_{r\times m}+\boldsymbol{E}_{n\times m} \tag{2}$$

其中，$\boldsymbol{E}_{n\times m}$ 为噪声矩阵。假设噪声服从不同的概率分布，通过最大似然函数，就可以得到不同类型的目标函数。

考虑噪声为高斯噪声，损失函数为

$$L(\boldsymbol{W},\boldsymbol{H})=\sum_{ij}\left[X_{ij}-(WH)_{ij}\right]^2 \tag{3}$$

该损失函数即为 2 范数损失函数。

若噪声为泊松噪声，损失函数为

$$L_{KL}(\boldsymbol{W},\boldsymbol{H})=\sum_{ij}\left[X_{ij}\,\mathrm{lb}\frac{X_{ij}}{(WH)_{ij}}-X_{ij}+(WH)_{ij}\right] \tag{4}$$

该损失函数即为 KL 离散度(KL 距离)。

为了消除尺度对于基矩阵的影响，对 $\boldsymbol{W}$ 限制其 1 范数为 1。对应于式(3)的乘性迭代规则为

基金项目：国家自然科学基金资助项目(60673090)；国家“973”计划基金资助项目(2005CB321800)
作者简介：翟亚利(1982－)，女，硕士研究生，主研方向：数据处理，信息检索；吴 翊，教授、博士生导师
收稿日期：2007-09-25 **E-mail：**laozei1216@sina.com

$$W_{ik} \leftarrow W_{ik} \frac{\left(\boldsymbol{XH}^{\mathrm{T}}\right)_{ik}}{\left(\boldsymbol{WHH}^{\mathrm{T}}\right)_{ik}}$$

$$W_{ik} \leftarrow \frac{W_{ik}}{\sum_l W_{lk}}$$

$$H_{kj} \leftarrow H_{kj} \frac{\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}\right)_{kj}}{\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{WH}\right)_{kj}}$$

同样对应于式(4)的乘性迭代规则为

$$W_{ik} \leftarrow W_{ik} \sum_j H_{kj} X_{ij} \Big/ (WH)_{ij}$$

$$W_{ik} \leftarrow \frac{W_{ik}}{\sum_l W_{lk}}$$

$$H_{kj} \leftarrow H_{kj} \sum_i W_{ik} X_{ij} \Big/ (WH)_{ij}$$

文献[1, 4]严格证明了该迭代算法的收敛性。乘性迭代规则严格保证了迭代过程中每一步结果的非负性。该迭代算法中一般采用随机初始化。

3 初始化方法

3.1 有监督主成分分析[5]

给定 m 个 n 维向量 $\boldsymbol{X}_i, i=1,2,\cdots,m$ ，假定数据已经中心化，即

$$\sum_{i=1}^{m} X_{i\alpha} = 0, \alpha = 1,2,\cdots,n$$

令

$$dist_{ij}^{p} = \sqrt{\sum_{\alpha=1}^{p} \left(X_{i\alpha}^{*} - X_{j\alpha}^{*}\right)^2}$$

表示在低维空间内 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 之间的欧氏距离。

定义单位 Laplacian 矩阵，用 $\boldsymbol{L}^u$ 表示：

$$L_{ij}^{u} = \delta_{ij} \cdot m - 1, \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

有文献对 PCA 作了如下的定理和引理：

定理 PCA 的 P 维投影方向就是保证下式最大化的投影方向：

$$\sum_{i<j} \left(dist_{ij}^{P}\right)^2$$

引理 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^u\boldsymbol{X} = m^2 \cdot \boldsymbol{S}$, $\boldsymbol{S} = \frac{1}{m}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 。

根据定理和引理的结论，在保证 $\boldsymbol{L}$ 为 Laplacian 矩阵的前提下，更换矩阵中的某些元素的大小，则

$$\arg\max\left(\sum_{i<j} d_{ij} \left(dist_{ij}^{P}\right)^2 \right)$$

的解为 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{d}}\boldsymbol{X}$ 的最大 P 项特征向量， $\boldsymbol{L}^{\mathrm{d}}$ 为 Laplacian 矩阵。

上式实质上表示了各点距离平方的加权和，称为加权 PCA(WPCA)，根据各量之间权值的大小来控制想要的各元之间的制约关系。假设 $\{d_{ij}\}_{i,j=1}^{m}$ 为对称非负权值，用其度量低维空间中数据点间的远离程度，当 $i=j$ 时， $d_{ij}=0$ 。此时$\boldsymbol{L}$为

$$L_{ij}^{d} = \begin{cases} \sum_{j=1}^{m} d_{ij} & i = j \\ -d_{ij} & i \neq j \end{cases}$$

SPCA 实际上是上述加权 PCA 的特殊情形，对于有监督模式识别，可以通过对类内的相异度乘以一个衰减因子 $t(0 \le t \le 1)$来增强类间的区分，修正的相异度为

$$\boldsymbol{d}_{ij}^{labled} = \begin{cases} t \cdot d_{ij} & i, j\text{同类} \\ d_{ij} & \text{否则} \end{cases}$$

一般情况下取 t 为 0，这意味着每一类的内部结构直接由元素的类别关系决定。

在实际应用中，SPCA 与 PCA 一样会遇到大矩阵分解问题，对于 PCA，可以通过对 $\boldsymbol{XX}^{\mathrm{T}}$ 进行谱分解得到 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的谱分解。类似地，需要寻求 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{d}}\boldsymbol{X}$ 的谱分解的优化算法，其中，$\boldsymbol{L}^{\mathrm{d}}$ 为 Laplacian 矩阵且主对角元素为正。

由矩阵的相关知识[6]有：对于非负定矩阵 $\boldsymbol{L}^{d}$ 可以通过谱分解得到 $\boldsymbol{L}^{\mathrm{d}} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ ，从而可以使用与PCA相同的方法计算 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{d}}\boldsymbol{X}$ 的谱分解。

对 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}^{\mathrm{d}}\boldsymbol{X}$ 进行谱分解之后，取最大 r 个特征值对应的特征向量为基矩阵 $\boldsymbol{W}$ 的初始化矩阵，进而根据最小二乘算法且令矩阵中负值元素为 0 得到 $\boldsymbol{H}$ 的初始化矩阵。

3.2 模糊C均值[7]

聚类分析的基本思想是用相似性尺度来衡量事物之间的亲疏程度，并以此来实现分类，模糊聚类分析的实质是根据研究对象本身的属性来构造模糊矩阵，在此基础上根据一定的隶属度确定其分类关系。

对于模糊聚类分析，设数据集 $\boldsymbol{X}$ 中含有 m 个样本，表示为 $X_k, k=1,2,\cdots,m$ ， c 为聚类数，设 $\boldsymbol{X}$ 中的任意样本 X_k 对第 i 类的隶属度为 u_{ik} ， $0 \le u_{ik} \le 1$。因此，其分类结果可用一个 $c \times m$ 阶矩阵 $\boldsymbol{U}$ 来表示，称其为模糊矩阵：

$$u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1, \forall k \text{ 及 } 0 < \sum_{k=1}^{m} u_{ik} < m, \forall i$$

聚类准则，目标函数 $J(\boldsymbol{U},V)$ 为

$$J(\boldsymbol{U},V) = \sum_{k=1}^{m}\sum_{i=1}^{c} (u_{ik})^{t} (d_{ik})^2$$

其中，$\boldsymbol{U}$ 为模糊矩阵； $V = \{V_1, V_2, \cdots, V_c\}$ 为 c 个聚类中心集合，$V_i \in R^n$ ； $t \in [1,\infty]$ 为加权指数； d_{ik} 为第 k 个样本到第 i 类的距离，定义为

$$(d_{ik})^2 = \|X_k - V_i\|^2 = (X_k - V_i)^{\mathrm{T}} \boldsymbol{A} (X_k - V_i)$$

其中，矩阵 $\boldsymbol{A}$ 为对称矩阵，当 $\boldsymbol{A} = \boldsymbol{I}$ 时，为欧氏距离。

使得目标函数最小化可推导出如下的迭代规则：

$$\begin{cases} u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{2}{t-1}} \right) & \text{当} I_k = \Phi \\ u_{ik} = 0, \forall i \in \tilde{I}_k \text{且} \sum_{i \in I_k} u_{ik} = 1 & \text{当} I_k \neq \Phi \end{cases} \tag{5}$$

$$V_i = \frac{1}{\sum_{k=1}^{m} (u_{ik})^{\mathrm{T}}} \sum_{k=1}^{m} (u_{ik})^{\mathrm{T}} X_k \quad i = 1,2,\cdots,c \tag{6}$$

其中，

$$\begin{cases} I_k = \{ i \mid 1 \le i \le c, d_{ik} = 0 \} \\ \tilde{I}_k = \{1,2,\cdots,c\} - I_k \end{cases}$$

若数据集 $\boldsymbol{X}$、聚类类别数 c 和权重 t 已知，就可通过式(5)和式(6)确定最佳的分类矩阵和聚类中心。算法结束后可设置分割门限 a (一般设为 0.5)，通过下式确定类别：

$u_{ik} = \max\{u_{1k}, u_{2k}, \cdots, u_{ck}\} \ge \alpha$ ，则 $X_k \in$ 第 i 类

该算法称为 FCM。本文选取 m 为 2(Matlab 自带语句中 m 的默认值)，作为初始化只要得到最佳的聚类中心即可。

以最佳的聚类中心为基矩阵 $\boldsymbol{W}$ 的初始化矩阵，进而与 SPCA 类似得到 $\boldsymbol{H}$ 的初始化矩阵。

4 实验介绍和结果分析

4.1 实验介绍

本文使用中国科学院提供的文本分类语料进行实验，从 http://www.nlp.org.cn/获得样本。基于 NMF 的文本分类方法的步骤为：

(1)分词。本文采用中科院计算所研制的汉语词法分析系统 ICTCLAS 进行分词处理，该分词系统分词正确率高达 97%

以上。

(2)特征选择。采用复合特征选择方法(将文档频率小于4的词条舍去)，使用改进的互信息方法进行特征选择。

互信息方法为

$$MI(t,c)=\mathrm{lb}\frac{p(t,c)}{p(t)\times p(c)}$$

改进的互信息特征选择方法为

$$MI\_N(t,c)=\mathrm{lb}\frac{p(t,c)}{p(t,\bar{c})\times p(c)}$$

其中，$p(t,c)$ ($p(t,\bar{c})$)：包含词条 t 属于(不属于)类别 c 的文档在语料中的概率；$p(t)$ ：包含词条 t 的文档在语料中的概率；$p(c)$ ：c 类文档在语料中的概率。若 $p(t,\bar{c})=0$ ，即 $p(t,c)=p(t)$ ，令 $p(t,\bar{c})=1/N$ ，该改进方法考察了词条对于类别的性能。

(3)计算词条权重。采用向量空间模型，使用常用的TFIDF公式计算选出特征的权重值：

$$w(t,d)=\frac{tf(t,d)\times \mathrm{lb}\left(N/n_t+0.01\right)}{\sqrt{\sum_{t\in d}\left[tf(t,d)\times \mathrm{lb}\left(N/n_t+0.01\right)\right]^2}}$$

其中，$w(t,d)$ 表示特征 t 在文档 d 中的权重；$tf(t,d)$ 表示特征 t 在文档 d 中出现的频数；N 为总的文本数；n_t 为出现特征 t 的文本数；0.01为加权常数。

(4)使用 NMF 进行降维处理。分别采用随机化、PCA、SPCA 和 FCM 来确定迭代算法的初始值。

(5)对于测试样本，按训练样本选出的词条为特征并根据权重公式计算对应的权重大小。进而，对于由训练样本得到的分解后的基矩阵 $\boldsymbol{W}$，在同样的基矩阵下，运用最小二乘算法并令矩阵中负值元素为0，得到测试样本对应的系数矩阵。

(6)对系数矩阵 $\boldsymbol{H}$，使用常用的 KNN 分类器进行检验，KNN 中近邻个数 K 取9，对于分类结果，采用常见的 $F1$ 值平均值进行度量。

记 a 为判断为正例的正例个数，b 为判断为正例的反例个数，c 为判断为反例的正例个数。则查全率为

$$recall=\frac{a}{a+c}$$

查准率为

$$precision=\frac{a}{a+b}$$

查全率与查准率是相互制约的，通常，提高查全率必然会降低查准率，反之亦然。因此要客观评估文本分类器的分类效果就必须同时考察这2个指标。本文采用 $F1$ 进行度量

$$F1=\frac{2recall\times precision}{recall+precision}$$

4.2 实验结果

实验中，按时间顺序从计算机、艺术、教育、交通、环境、军事、医药7类文本集中分别选择100篇文本作为训练集，分别选择 30 篇作为测试集进行实验。分词后得到词条138 855个，复合特征选择后选取2 000个特征，并用TFIDF公式计算得到词条文本矩阵，即 $2\,000\times 710$ 的数据集 X。

实验1：使用随机初始化，确定最优迭代次数

取定维数 r 为150，分别以式(3)和式(4)为损失函数，计算不同迭代次数对应的损失函数值。其结果如图1和图2所示。由2幅图综合可知：在以后的实验中对于2种迭代算法都取迭代次数为30。

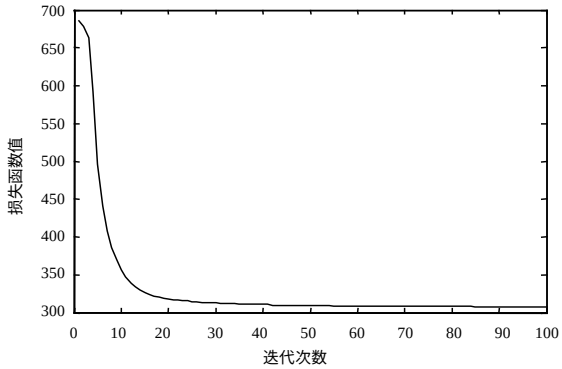

图1　以欧氏距离为损失函数的迭代次数-损失值图

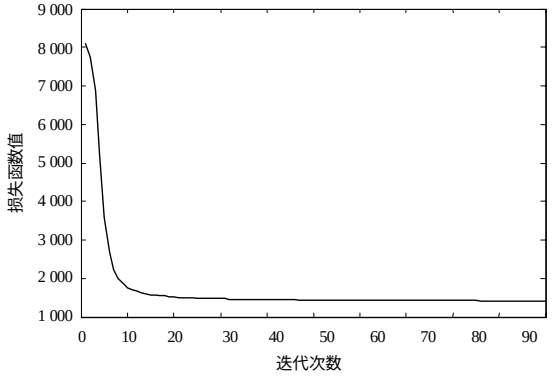

图2　以KL距离为损失函数的迭代次数-损失值图

实验2：r 不断变化时，4种初始化方法的分类情况

由于随机初始化的分类结果受初始化影响比较大，因此在图3、图4中没有画出随机初始化对应的结果曲线。

基矩阵的维数 r 的取法为：在150之前以25为间隔，200~500以50为间隔。对应于式(3)和式(4)的损失函数，其实验结果分别如图3和图4所示。

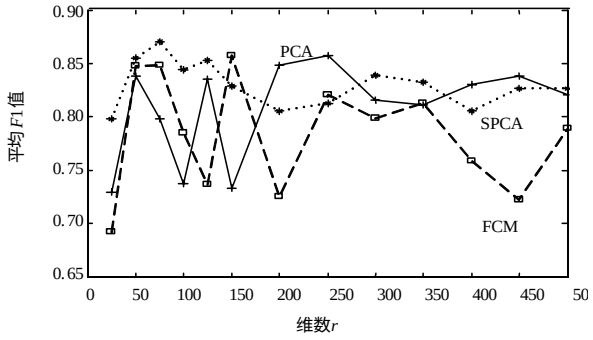

图3　以欧氏距离为损失函数的维数-平均 $F1$ 值图

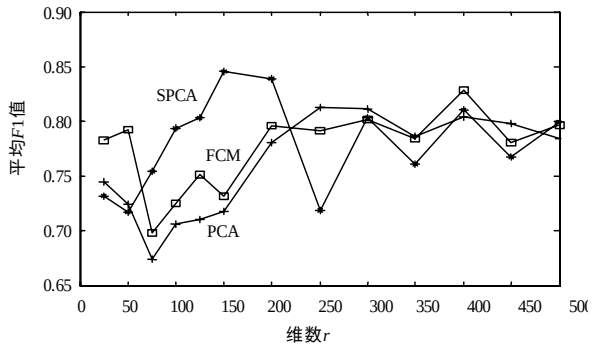

图4　以KL距离为损失函数的维数-平均 $F1$ 值图

(下转第197页)