◎数据库与信息处理◎

基于非负矩阵分解的特征向量抽取方法特点研究

郭 勇[1,2],鲍丽春[3]
GUO Yong[1,2],BAO Li-chun[3]

1.国防科技大学 信息系统与管理学院,长沙 410073
2.北京系统工程研究所,北京 100101
3.北京航天飞控中心,北京 100094
1.Information System and Management College,National University of Defense Technology,Changsha 410073,China
2.Beijing Institute of System Engineering,Beijing 100101,China
3.Beijing Aerospace Control Centre,Beijing 100094,China

GUO Yong,BAO Li-chun.Characteristics of non-negative matrix factorization for feature extraction.Computer Engineering and Applications,2007,43(17):154-156.

Abstract: Non-negative Matrix Factorization(NMF) is a new algorithm for feature extraction.This paper compares NMF with three other existing feature extraction method:Singular Value Decomposition(SVD) is fundamentally different from NMF in feature extraction,but the non-negative constraints make the decomposition procedure of NMF much more like the process of human cognition than SVD;the clustering-based feature extraction method can be considered as a simplified NMF algorithm;and the probabilistic-based feature extraction method is proved to be one type of NMF algorithm with special constraints.Through these comparisons,we catch the non-negative and local characteristics of NMF.

Key words: feature extraction;non-negatives matrix factorization;feature vector

摘 要:非负矩阵分解算法可以作为一种新型的特征抽取方法。将非负矩阵分解算法和现有的其它三种现有的特征抽取算法进行详细比较:奇异值分解方法和非负矩阵分解方法本质上是不同的两种特征抽取方法,非负特性使得由非负矩阵分解比奇异值分解方法更接近人们的认知习惯。基于聚类的特征提取方法是一种简化了的非负矩阵分解算法;基于概率的特征提取方法等价于非负矩阵分解在特定约束条件下的变体。通过比较充分体现了非负矩阵分解算法的非负性和局部性特点。

关键词:特征抽取;非负矩阵分解;特征向量
文章编号:1002-8331(2007)17-0154-03 **文献标识码**:A **中图分类号**:TP301.6

1 引言

随着科学技术的发展,生成和收集信息的能力显著提高,急需一些智能工具来辅助检索、处理这些信息。作为信息处理中的一项核心技术,特征抽取一直受到广泛重视。特征抽取研究内容包括特征构建和选择、空间降维、稀疏表示等等,它将被运用于生物信息、化学分析、文本处理、模式识别、语音和图像处理等等众多领域,具有重要的研究意义。

最近出现了一种新的算法——非负矩阵分解(NMF,Non-negative Matrix Factorization)算法[1]。NMF 将非负样本矩阵 $\boldsymbol{X}=(x_{ij})_{m\times n}$ 分解为一个 $m\times r$ 维非负矩阵 $\boldsymbol{U}=(u_{ij})_{m\times r}$ 和一个 $r\times n$ 维非负矩阵 $\boldsymbol{V}=(v_{ij})_{r\times n}$ 的乘积,并满足 $\boldsymbol{X}\approx\boldsymbol{UV}$。为了达到好的分解效果,NMF 在分解结果中尽可能体现样本的主要信息,因此 NMF 为特征抽取提供了一条新的途径。NMF 算法在特征抽取领域的一个成功应用是对人脸图像特征的抽取[1]:通过分解特定条件下一组人脸部图像(每幅图像形成一个向量)得到预定数目的基向量。而每个基向量(一幅图像)显示的正是诸如“鼻子”、“嘴巴”、“眼睛”等人脸局部特征信息。原图像表示为这些局部特征的加权组合。之后众多研究将 NMF 及其变体算法作为特征抽取算法,运用于图像[2,3]、文本[4-6]、日志分析等领域[7-9],均取得了很好的效果。

在 NMF 之前,已经存在多种特征提取算法,包奇异值分解算法[10]、基于聚类的特征抽取方法[11],以及基于概率的特征抽取算法[12]。NMF 跟这三个算法相比有什么特点和关系?本文将从抽取特征信息的角度,进行详细比较分析。

2 非负矩阵分解和奇异值分解的比较

奇异值分解(SVD,Singular Value Decomposition)算法将矩阵 $\boldsymbol{X}$ 分解为三个矩阵 $\boldsymbol{U}=(u_{ij})_{m\times m}$,$\boldsymbol{\Sigma}=(\Sigma_{ij})_{m\times n}$,$\boldsymbol{V}=(v_{ij})_{n\times n}$,使得 $\boldsymbol{X}=\boldsymbol{U}\sum\boldsymbol{V}^{\mathrm{T}}$,并满足 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}_m$,$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}_n$,式中 $\boldsymbol{\Sigma}=diag(\sigma_1,\sigma_2,\cdots,\sigma_p)$,$p$ 表示概念语义数目,通常远小于 $\min(m,n)$。矩阵 $\boldsymbol{\Sigma}$ 中的元素被称为奇异值,并按降序排列,即 $\sigma_1\geqslant\sigma_2\geqslant\cdots\geqslant\sigma_p\geqslant 0$。好的近似分解结果是通过选择较大奇异值和与之对应的 $\boldsymbol{U}$ 中的列向量和 $\boldsymbol{V}$ 中的行向量构成的。被选中的 $\boldsymbol{U}$ 中的列向量就被认为是抽取得到的特征向量。SVD 和 NMF 算法在以下几个方面存在不同:

2.1 问题本质

从本质上讲,NMF 的分解过程是一个带约束的非线性规划问题,通过规划使得分解结果 $\boldsymbol{UV}$ 和样本数据 $\boldsymbol{X}$ 尽可能接近。SVD 的分解过程的核心问题是求解矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的特征值和特征向量,SVD 能让分解结果 $\boldsymbol{U\Sigma V}^{\mathrm{T}}$ 和样本数据 $\boldsymbol{X}$ 完全相等。

NMF 所对应的非线性规划很难得到最优解，只能从初始条件出发得到局部最优解，同时 NMF 方法的分解结果受初始值影响很大,这是 NMF 的缺点。而 SVD 则相反,SVD 在分解过程有确定的方法和步骤，对相同的矩阵 $\boldsymbol{X}$,SVD 方法一般都能够得到确定的分解结果。

2.2　抽取特征向量的特点

一般把 NMF 的分解结果 $\boldsymbol{UV}$ 中矩阵 $\boldsymbol{U}$ 的列向量看作是由 NMF 抽取的特征向量，把 SVD 的分解结果 $\boldsymbol{U\Sigma V}^{\mathrm{T}}$ 中 $\boldsymbol{U}$ 矩阵的列向量看作是 SVD 抽取的特征向量。

从形式上看 NMF 和 SVD 抽取的特征向量存在很大差别。NMF 抽取的特征向量具有非负的特点，这使得特征向量之间的内积均大于零,因此 NMF 抽取的特征向量不可能完全正交,这必然会带来信息冗余。而 SVD 能保证抽取的特征向量两两正交,但是它们失去了非负的特点。

从特征向量对样本的表达能力来看，两者也有很大不同。关于 SVD 有如下定理：假设矩阵 $\boldsymbol{X}$ 的分解结果 $\boldsymbol{X=U\Sigma V}^{\mathrm{T}}$,用 $\boldsymbol{\Sigma}_k=diag(\sigma_1,\sigma_2,\cdots,\sigma_k)$ 表示 $\boldsymbol{\Sigma}$ 的中最大的 k 个奇异值构成的对角阵,$\boldsymbol{U}_k=(u_1,u_2,\cdots,u_k)$,$\boldsymbol{V}_k=(v_1,v_2,\cdots,v_k)$,分别代表 $\boldsymbol{\Sigma}_k$ 所对应 $\boldsymbol{U}$、$\boldsymbol{V}$ 的子矩阵,那么可以证明:在所有秩为 k 的矩阵 $\boldsymbol{C}$ 中 $\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}}$ 是对 $\boldsymbol{X}$ 的最好的近似,即当 $\boldsymbol{C=U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}}$ 时候，$\|\boldsymbol{X-C}\|=\sum_{ij}(\boldsymbol{X}_{ij}-\boldsymbol{C}_{ij})^2$ 达到最小值[13]。这就是说 SVD 抽取的特征向量在不同的维度都能够对样本向量作最好的近似,使得近似后的结果和样本的误差在 L_2 距离定义下达到最小。NMF 没有这个特点,首先 NMF 在分解过程中基向量的数目是确定的,每次分解只能得到在一个确定维度下对样本进行近似。其次 NMF 作为规划问题，虽然它规划的目标是让 $\boldsymbol{X}$ 和 $\boldsymbol{UV}$ 之间的误差最小,但是它只能保证规划的结果在局部达到最优，并不能像 SVD 那样保证误差的全局最小。

从特征向量之间的区别来看,SVD 和 NMF 也是不同的。SVD 所抽取的特征向量在体现样本特征方面是有很大差别的。假设基向量 $\boldsymbol{u}_i$ 对应最大的奇异值 σ_i，那么样本在 $\boldsymbol{u}_i$ 上的投影能量总和为 $\boldsymbol{E}_i=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}_j)^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}_j)=\sigma_i^2$。这样对应奇异值越大的向量具有更大的样本投影能量,在体现了样本的特性方面显得更重要。而 NMF 所抽取的特征向量间重要程度差别不大,每个特征向量都用于表达样本的某一部分特征信息。

2.3　特征向量的解释

对矩阵分解算法加入非负限制的原因之一就是可以对所抽取的特征向量赋予更好的解释。由 NMF 抽取的特征向量,可以给出很好的物理含义。比如在图像领域,一个非负特征向量也可以被解释为一幅特征图,向量中的每个元素代表图中相应点的像素值。在文本领域,一个非负特征向量可以被解释为一个“主题”，向量中每个元素代表某个单词在主题中的重要程度。而对 SVD 抽取的特征向量,由于没有非负限制,可能会出现负值,因此很难给出很好的物理解释。

总的来说,SVD 仅仅从代数角度来分析样本矩阵，往往只能抽取样本的代数特征。虽然从数学角度看,SVD 能获的分解结果比较完美且易于分析，但是负数的存在使 SVD 的分解结果失去了对实际问题的联系。而 NMF 加入了非负限制,虽然分解很困难,分解结果也不是数学上的最优解,但是非负限制使得问题的处理更符合人的思维特性,分解结果更容易解释。

3　非负矩阵分解和 k–均值算法的比较

聚类算法也可用于抽取特征向量。基于聚类的特征抽取方法的基本思想是通过聚类算法将相似的样本聚成一类,聚类中心向量即被认为是特征向量。将 NMF 和经典的 k–均值算法[14]进行比较。对于给定的 n 个样本 $\{x_i\}_{i=1}^n$ 和常数 k,k–均值算法将样本聚为 k 个类 $\{\pi_j\}_{j=1}^k$,对应可以获取 k 个聚类中心向量 $\{c_j\}_{j=1}^k$。和 NMF 算法相比,聚类算法的优点是简单、处理速度快。但是由聚类中心来表示生成的特征向量之间会存在信息冗余。下面从几个方面对 k–均值算法和 NMF 算法得到的特征向量作详细的对比。

3.1　目标函数

k–均值算法试图从给定的 n 个样本 $\{x_i\}_{i=1}^n$ 中抽取 k 个特征向量 $\{\boldsymbol{u}_j\}_{j=1}^n$,使得目标函数 Q_1 达到最小值：

$$Q_1(\boldsymbol{X},\boldsymbol{U})=\sum_{i=1}^{n}\min_{j=1}^{k}\|u_j-x_i\| \tag{1}$$

从特征抽取的角度来看,NMF 算法的分解过程也可以理解为:试图从给定的 n 个样本 $\{x_i\}_{i=1}^n$ 中抽取 k 个特征向量 $\{\boldsymbol{u}_j\}_{j=1}^k$,使得目标函数 Q_2 达到最小值：

$$Q_2(\boldsymbol{X},\boldsymbol{U})=\sum_{i=1}^{n}\min_{\Delta v}\|Uv-\boldsymbol{x}_i\| \tag{2}$$

式中 $\boldsymbol{U}=(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,u_k)$,$\boldsymbol{v}$ 是一个元素均非负的 $\boldsymbol{k}$ 维的列向量。

两种方法从目标函数上看都试图从给定的 n 个样本 $\{x_i\}_{i=1}^n$ 中抽取 k 个特征向量 $\{\boldsymbol{u}_j\}_{j=1}^k$。且两者目标函数 Q_1 和 Q_2 形式上相近,不同的是 Q_1 含义为:用 $\{\boldsymbol{u}_j\}_{j=1}^k$ 中的单个向量作为样本近似,达到误差最小。Q_2 含义为用 $\{\boldsymbol{u}_j\}_{j=1}^k$ 中向量的非负线性组合作为样本近似,达到误差最小。从这个意义上看,Q_1 是 Q_2 的特例,当 Q_2 中 Δv 的选择范围被约束为单位基向量时,Q_2 就退化为 Q_1。

3.2　特征向量的取值范围

基于聚类的特征向量抽取方法将各个类中心向量作为特征向量,提取的特征向量均是一个类的中心向量。由此可知,用 k–均值聚类算法抽取特征向量时,特征向量的选取范围是依赖于样本向量的,它只能是某部分样本的中心,因此其选择范围是有限的—对于 n 个样本,通过聚类算法所得到的中心向量必属于一个含有 2^n 个元素的集合中。

而 NMF 算法在抽取特征向量时，其选择范围是不依赖于样本的,它可以在一个与样本维数相同的非负连续空间中任意选取,选取范围远远大于 k–均值算法。可见聚类算法的特征向量选取范围仅仅是 NMF 算法的一个有限子集。

3.3　特征向量的局部性特点

用 k–均值算法抽取的每个特征向量都是一个样本集合的中心。同时由目标函数可知,用这些特征向量本身来近似代替原样本,误差是最小的。因此用 k–均值算法抽取的特征向量本身和样本是很接近的,反应了样本的整体特征。

与之相反,用 NMF 算法抽取的 k 个特征向量 $\{\boldsymbol{u}_j\}_{j=1}^k$ 和样本不一定非常接近,但是这些特征向量的非负线性组合能达到对样本最佳近似效果，因此 NMF 算法抽取的特征向量往往和

样本不是很接近,但是能体现样本的局部特征。

总的来说,从目标函数和特征向量的选取范围可知,k-均值算法是NMF算法经过简化后的特例,当NMF目标函数中$\boldsymbol{U}$的各列向量的取值范围被限制为一个仅包含2^n个元素的样本中心向量集合,并且$\boldsymbol{V}$中各列向量取值被限制为单位基向量时,NMF算法就退化为k-均值算法。同时这样的简化使得NMF失去了可抽取局部特征的特点,只能抽样本的整体特征。

4 非负矩阵分解和概率潜在语义索引的比较

在文本领域存在一种概率潜在语义索引(PLSI,Probability Latent Semantic Indexing)模型,它从概率角度分析单词、文本和文本特征(主题)三者之间的关系。该模型假设人们对任何问题的描述都是围绕某个主题展开的。各个主题概念之间具有相对明显的界线(可以认为是相互独立的)。

设文本集为$D=[d_1,d_2,\cdots,d_n]$,词汇集为$T=[t_1,t_2,\cdots,t_m]$,且文本的主题集为$Z=[z_1,z_2,\cdots,z_k]$,文本$d\in D$的产生模型表述为:(1)以一定的概率$p(d_j)$选择文本d_j;(2)任意文本d_j含有主题z_k的概率为$p(z_k|d_j)$;(3)在主题z_k的条件下,产生词t_i的概率为$p(t_i|z_k)$。

经过上述过程获得观测点对(d,t),潜在的主题z被忽略掉,产生下述联合概率模型:

$$p(t_i,d_j)=p(d_j)p(t_i|d_j)=p(d_j)\sum_{l=1}^{k}p(t_i|z_l)p(z_l|d_j) \tag{3}$$

其对应的最大似然估计表达式如下:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}n(t_i,d_j)\log p(t_i,d_j) \tag{4}$$

式中$n(t,d)$表示词t在文本d中出现的次数。通过EM算法可以求得参数$p(t_i|z_k)$和$p(z_k|d_j)$的值。

PLSI从概率角度出发建立模型,所以能给分解结果更好的概率解释,但是它的求解过程却和NMF很相似,可以证明它等价于加入限制后的一种NMF算法变体。

PLSI通过EM方法,求参数$p(t_i|z_k)$和$p(z_k|d_j)$的值,使得最大似然估计表达式$\sum_{i=1}^{m}\sum_{j=1}^{n}n(t_i,d_j)\log p(t_i,d_j)$达到最大。这从本质上讲,也是一个规划问题。如果构造参数矩阵$\boldsymbol{U}$和$\boldsymbol{V}$,使得$\boldsymbol{U}_{ik}=p(t_i|z_k)$,$\boldsymbol{V}_{kj}=p(z_k|d_j)$。则该问题就对应为如下的规划问题:

$$\begin{cases}\underset{\boldsymbol{U},\boldsymbol{V}}{\text{Max}}\sum_{i=1}^{m}\sum_{j=1}^{n}n(t_i,d_j)\log p(t_i,d_j)\\ \boldsymbol{U}\geqslant 0,\boldsymbol{V}\geqslant 0\\ \sum_{i=1}^{m}\boldsymbol{U}_{ij}=1,j=1,\cdots,k\\ \sum_{i=1}^{k}\boldsymbol{V}_{ij}=1,j=1,\cdots,n\end{cases} \tag{5}$$

同时,对式(5)中的目标函数进行化简。由于:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}n(t_i,d_j)\log p(t_i,d_j)=$$

$$\sum_{j=1}^{n}n(d_j)\log P(d_j)+\sum_{i=1}^{m}\sum_{j=1}^{n}n(t_i,d_j)\log\sum_{l=1}^{k}P(t_i|z_l)P(z_l|d_j)$$

公式中$n(t_i,d_j)$、$p(d_j)$均为常数,$n(d_j)$代表文本d_j所含的单词总数,也为常数。因此规划问题(5)中的目标函数就可以化简为:

$$\underset{\boldsymbol{U},\boldsymbol{V}}{\text{Max}}\sum_{i=1}^{m}\sum_{j=1}^{n}n(t_i,d_j)\log(\boldsymbol{UV})_{ij} \tag{6}$$

再来考虑NMF,如果构造矩阵$\boldsymbol{X}$,使得$\boldsymbol{X}_{ij}=n(t_i,d_j)$;同时取$D(\boldsymbol{X}\parallel\boldsymbol{UV})=\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}\log\frac{\boldsymbol{X}_{ij}}{[\boldsymbol{UV}]_{ij}}-\boldsymbol{X}_{ij}+[\boldsymbol{UV}]_{ij})$为距离函数,进行非负矩阵分解得结果$\boldsymbol{UV}$,使$\boldsymbol{X}\approx\boldsymbol{UV}$。并在原有的非负约束条件基础上,规定分解结果中矩阵$\boldsymbol{U}$,$\boldsymbol{V}$须满足$\sum_{i=1}^{n}\boldsymbol{U}_{ij}=1,j=1,2,\cdots,k$,$\sum_{i=1}^{k}\boldsymbol{V}_{ij}=1,j=1,2,\cdots,m$。这样,该非负矩阵分解问题就对应为如下的规划问题:

$$\begin{cases}\underset{\boldsymbol{U},\boldsymbol{V}}{\text{Min}}\sum_{i=1}^{m}\sum_{j=1}^{n}(\boldsymbol{X}_{ij}\log\frac{\boldsymbol{X}_{ij}}{[\boldsymbol{UV}]_{ij}}-X_{ij}+[\boldsymbol{UV}]_{ij})\\ \boldsymbol{U}\geqslant 0,\boldsymbol{V}\geqslant 0\\ \sum_{i=1}^{m}\boldsymbol{U}_{ij}=1,j=1,\cdots,k\\ \sum_{i=1}^{k}\boldsymbol{V}_{ij}=1,j=1,\cdots,n\end{cases} \tag{7}$$

式中$\boldsymbol{X}_{ij}=n(t_i,d_j)$是常数,又因为$\sum_{i=1}^{m}\boldsymbol{U}_{ij}=1$,$\sum_{i=1}^{k}\boldsymbol{V}_{ij}=1$,故$\sum_{i=1}^{m}\sum_{j=1}^{n}[\boldsymbol{UV}]_{ij}=n$,也为常数,因此规划问题(7)的目标函数即可化简为:

$$\underset{\boldsymbol{U},\boldsymbol{V}}{\text{Max}}\sum_{i=1}^{m}\sum_{j=1}^{n}\boldsymbol{X}_{ij}\log(\boldsymbol{UV})_{ij} \tag{8}$$

跟式(6)完全等价。由此可知式(5)和式(7)这两个规划问题是全等的。

可以说从分解过程来考虑,PLSI即是NMF一种变体,该变体对NMF的分解结果加入归一化限制$\sum_{i=1}^{m}\boldsymbol{U}_{ij}=1$,$\sum_{i=1}^{k}\boldsymbol{V}_{ij}=1$。这样的归一化限制使得PLSI的分解结果更具概率意义,而传统的EM方法也成了一种特殊矩阵的分解方法。

5 结论

本文介绍了基于NMF的特征向量抽取方法,将NMF和现有的三种典型的特征抽取算法—SVD分解算法和基于聚类的特征抽取算法、基于概率的特征抽取算法—进行了详细比较。比较中发现SVD算法和NMF算法本质上差别很大,很多特性构成互补关系。基于聚类的特征抽取算法可以看成是一种简化了的非负矩阵分解算法,基于概率的特征抽取算法是一种加了归一化限制的非负矩阵分解算法。(收稿日期:2006年10月)

参考文献:

[1] Lee D D,Seung H S.Learning the parts of objects by non-negative matrix factorization[J].Nature,1999,401:788-791.

[2] Guillamet D,Vitrià J.Non-negative matrix factorization for face recognition[C]//The Catalonian Conference on AI:Topics in Artificial Intelligence,Castellón,Spain,2002:336-344.

[3] Li S,Hou X,Zhang H.Learning spatially localized,parts-based representation[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2001,Hawaii,USA,2001:207-212.

[4] Xu Bao-wen,Lu Jian-jiang,Huang Gang-shi.A constrained non-negative matrix factorization in information retrieval[C]//The 2003

(下转159页)