

一种基于 p-Laplacian 的谱聚类维数约简算法*

石远超, 马 宏, 李海涛

(国家数字交换系统工程技术研究中心, 郑州 450002)

摘要:近年来,谱聚类因其深厚的理论基础而在机器学习和数据挖掘领域中引起了广泛的关注。针对谱聚类算法中采用 Laplacian 矩阵时无法获得较好的图切判据,通过引入 p-Laplacian 算子,提出了一种基于 p-Laplacian 的谱聚类维数约简算法。仿真实验表明,提出的方法可以获得一种优化图切的近似解,使得在降维后能够更为精确地求得原始数据在低维空间中的嵌入投影。

关键词:谱聚类;图切判据;维数约简;p-Laplacian

中图分类号: TP18 **文献标志码:** A **文章编号:** 1001-3695(2012)06-2126-03

doi:10.3969/j.issn.1001-3695.2012.06.032

Dimensionality reduction algorithm based on p-Laplacian in spectral clustering

SHI Yuan-chao, MA Hong, LI Hai-tao

(China National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China)

Abstract: In recent years, spectral analysis approaches have received much attention in machine learning and data mining areas, due to their rich theoretical foundations. This paper addressed the problem of Laplacian in spectral clustering, which couldn't get ideally graph cut criterion, by introducing the p-Laplacian operator, this paper proposed a new dimensionality reduction algorithm based on p-Laplacian. The experimental results denote that, the approach can get a approximation of the optimal graph cut, and can accurately get embedding mapped of original data in low dimensionality space.

Key words: spectral clustering; graph cut criterion; dimensionality reduction; p-Laplacian

0 引言

为了解决维数灾难的问题,维数约简方法的研究引起了研究者的普遍关注。维数约简算法主要分为线性和非线性方法两大类。线性方法主要包括主成分分析(PCA)、局部线性投影(LLP)、线性判别分析(LDA)和多维尺度变换(MDS)等。线性方法实现简单,但只能提取数据的线性部分,对嵌入在一个低维流形的非线性高维数据是无效的,而非线性方法可以很好地解决这个问题。非线性方法主要包括局部线性嵌入(LLE)、等距离映射(ISOMAP)、拉普拉斯特征映射(LE)等方法^[1],还可以利用谱聚类来构造非线性维数约简算法。因为谱聚类是通过相似性矩阵的特征向量来得到数据点在低维空间的嵌入,然后用经典聚类方法将嵌入后的数据点进行分组的一种方法,而它的低维嵌入是建立在特征分解的基础上,最终能得到一个更能表示原始数据的低维空间,从而能很好地实现非线性维数约简。

谱聚类算法中,合适的图切判据可以在低维嵌入后最大化类间散度,并同时最小化类内散度,从而获得较好的降维效果。针对经典谱聚类构建图切判据时,采用的拉普拉斯矩阵难以得到较好的图切判据,本文通过引入 p-Laplacian 算子来获得一种优化图切的近似解,并提出基于 p-Laplacian 的谱聚类维数约简算法(reduction algorithm based on p-Laplacian in spectral clustering, PLSC)。实验结果证明,该算法可以使投影到变换后的子空间中类间的数据点具有更好的区分度。

1 基本概念

1.1 谱图划分理论

谱聚类的思想来源于谱图划分理论。Donath 和 Hoffman 将图的划分问题与相似矩阵有关的特征向量联系起来,而 Fiedler 证明了图的二分法与 Laplacian 矩阵的第二特征向量密切相关^[2]。根据 Rayleigh-Ritz 理论,谱聚类算法求得的拉普拉斯矩阵的若干特征向量就是最优划分问题松弛后的近似解。由此可见,图的划分和 Laplacian 矩阵都是谱聚类算法的重要组成部分。

图切判据旨在寻求一种类内相似度尽量大,类间相似度尽量小的最优划分,同时兼顾了划分的均衡性。假设将数据点看做图中的一个顶点 V ,将两点之间的相似度作为边的权重 W ,就得到一个基于相似度的无向加权图: $G = (V, E)$,其中, V 是顶点的集合, E 是边的集合。聚类问题就相应地转换为在图 G 上的划分问题,其数学描述为

$$\text{cut}(A_1, \dots, A_k) = \sum_{i=1}^k \text{cut}(A_i, \overline{A_i}) \quad (1)$$

其中: $\text{cut}(A, B) = \sum_{i \in A, j \in B} W_{ij}$, $\overline{A_i} = \{v_p | v_p \in V \text{ 且 } v_p \notin A_i\}$ 是子图 A_i 中包含的顶点数目。该目标函数的最小化比较简单,但是当数据集存在离群点时,易产生倾斜划分。为了产生均匀的簇, Hagen 和 Kahng 提出了比率切准则,简称 Rcut^[2],定义为

$$\text{Rcut}(A_1, \dots, A_k) = \sum_{i=1}^k \frac{\text{cut}(A_i, \overline{A_i})}{|A_i|} \quad (2)$$

该目标函数对倾斜划分有一定的约束作用,但顶点数多并

收稿日期: 2011-11-26; 修回日期: 2011-12-30 基金项目: 国家“863”计划资助项目(2011AA010603)

作者简介: 石远超(1985-),男,河南濮阳人,硕士研究生,主要研究方向为通信与信息系统(relsys@163.com);马宏(1968-),男,副教授,主要研究方向为电信网安全;李海涛(1982-),男,讲师,主要研究方向为电信网安全。

非类内相似度大的充要条件,因此 Shi 和 Malik 提出了规范切,简称 Neut^[2],定义为

$$\text{Neut}(A_1, \dots, A_k) = \sum_{i=1}^k \frac{\text{cut}(A_i, \bar{A}_i)}{\text{vol}(A_i)} \quad (3)$$

其中: $\text{vol}(A_i) = \sum_{p \in A_i, q \in V} W_{pq}$ 。该目标函数要求子图间边权重的最小化,且子图内边权重的最大化,划分效果较好,但由于它是投影矩阵的非线性函数,算法可能存在局部极小。

在规范切的基础上,Cheeger 等人提出了更为优化的齐格切,简称 Ccut^[3],定义为

$$\text{Ccut}(A_1, \dots, A_k) = \sum_{i=1}^k \frac{\text{cut}(A_i, \bar{A}_i)}{\min(A_i)} \quad (4)$$

该目标函数可使类内的数据更紧凑,即某一类在聚类后所占的空间规模与该类中数据点规模的相关性尽量小,但根据 Rayleigh 商原理可知,计算最优的齐格切是一个 NP 难问题。下面通过在谱聚类中引入 p-Laplacian 来得到齐格切的一个近似解。

1.2 p-Laplacian 算子

谱图理论分析的基础是图的拉普拉斯矩阵。根据 Rayleigh 商原理,求解图切优化的问题就等同于求解拉普拉斯矩阵的特征值问题,它代表了最佳图划分的一个解。文献[4]中 Hein 等人给出了图的标准 Laplacian 算子 Δ_2 的内积形式的定义,表示为

$$\langle f, \Delta_2 f \rangle_i = \frac{1}{2} \sum_{j=1}^n w_{ij} (f_i - f_j)^2 \quad (5)$$

其中: f 是 Laplacian 的特征向量。若将式(5)一般化,设存在一个参数 $p \in (1, 2]$, 将 Laplacian 算子转换为 Δ_p , 则式(5)可表示为

$$\langle f, \Delta_p f \rangle_i = \frac{1}{2} \sum_{j=1}^n w_{ij} (f_i - f_j)^p \quad (6)$$

Laplacian 的矩阵形式 L 分为非规范和规范的形式。非规范的 Laplacian 矩阵表示为 $L = D - W$, 规范的形式表示为 $L = I - D^{-1}W$, 其中, W 是亲和度矩阵 (affinity matrix), D 为对角阵。由此不难得出相应的基于 p-Laplacian 的非规范矩阵和规范矩阵形式为

$$(\Delta_p^{(u)} f)_i = \sum_{j \in V} w_{ij} \varphi_p(f_i - f_j) \quad (7)$$

$$(\Delta_p^{(n)} f)_i = \frac{1}{d_i} \sum_{j \in V} w_{ij} \varphi_p(f_i - f_j) \quad (8)$$

其中: $d_i = \sum_{j=1}^n w_{ij}$ 为图的度函数 (degree function), $\varphi_p(x) = |x|^{p-1} \text{sign}(x)$, 且当 $p=2$ 时, $\varphi_2(x) = x$ 。为引入 p-Laplacian 的特征值 λ_p , 以非规范的 p-Laplacian 矩阵为例, 给出以下定义。

定义 1 如果存在一个实数 λ_p 满足式(9), 则 λ_p 为特征向量 f 所对应的特征值^[1]。

$$(\Delta_p^{(u)} f)_i = \lambda_p \varphi_p(f_i), \forall i = 1, \dots, n \quad (9)$$

由于在矩阵运算中, 只有那些对应于最小特征值的特征向量被认为含有重要辨别信息, 使得这一属性在聚类算法中有着极其重要的作用。相关研究证明, 在线性运算中, 如果 λ 是一个算子 Δ 的最小特征值, 应该满足使式(10)达到下界值为

$$\lambda = \arg \min_{f \in \mathbb{R}^n} \frac{\langle f, \Delta f \rangle}{\|f\|^2} \quad (10)$$

其中: $\|f\|^2 = \sum_{i=1}^n |f_i|^p$ 。如果把这一特性应用到非线性中的 p-Laplacian 中, 则可得到定义 2。

定义 2 如果 λ_p 是一个算子 Δ_p 的最小特征值, 应该满足使式(10)达到下界值^[4]为

$$\lambda_p = \arg \min_{f \in \mathbb{R}^n} \frac{\langle f, \Delta_p f \rangle}{\|f\|^p} \quad (11)$$

结合式(11), 针对非规范的 p-Laplacian, 可以构造一个目标函数 $F_p(f)$ 为

$$F_p(f) = \frac{\langle f, \Delta_p^{(u)} f \rangle}{\|f\|^p} = \frac{\frac{1}{2} \sum_{i,j=1}^n w_{ij} |f_i - f_j|^p}{\|f\|^p} \quad (12)$$

由上面的分析可知, 如果 λ_p 是一个算子 Δ 的最小特征值, 且 f 是 λ_p 对应的特征向量, 则应该满足使 $F_p(f)$ 达到下界边值, 该边值即所需要的特征值的大小, 进而获得那些对聚类比较重要的特征向量。同样, 可对规范的 p-Laplacian 构造函数 $F_p(f)$, 在此不再详述。本文将以非规范的 p-Laplacian 为例, 构建谱聚类维数约简算法。

2 改进的谱聚类维数约简算法

2.1 p-Laplacian 与齐格切的关系

由于齐格切可使类内的数据更紧凑, 即某一类在聚类后所占的空间规模与该类中数据点规模的相关性尽量小, 下面将通过分析引入的 p-Laplacian 算子与齐格切的关系, 从而通过 p-Laplacian 算子来得到一个齐格切的近似解。

对于一个分割问题 $(A_i, \bar{A}_i)$, 齐格切给出的目标函数为 $\text{Ccut}(A_1, \dots, A_k) = \sum_{i=1}^k \frac{\text{cut}(A_i, \bar{A}_i)}{\min(A_i)}$ 。这里首先定义一个关联分割 (f, A_i) , 表示为

$$(f, A)_i = \begin{cases} \frac{1}{|A_i|^{p-1}} & i \in A_i, p \in (1, 2] \\ \frac{1}{|\bar{A}_i|^{p-1}} & i \in \bar{A}_i, p \in (1, 2] \end{cases} \quad (13)$$

为了得到目标函数的最小特征值, 利用式(13)对式(12)进行变换

$$F_p(f, A_i) = \frac{\sum_{ij} w_{ij} \left| \frac{1}{|A_i|^{p-1}} - \frac{1}{|\bar{A}_i|^{p-1}} \right|^p}{\|f, A\|_p^p} \times 6 \quad (14)$$

其中: $\|f, A\|_p^p = \sum_i |f_i, A|^p = \frac{1}{|A_i|^{p-1}} - \frac{1}{|\bar{A}_i|^{p-1}}$, 则式(14)可转换为

$$F_p(f, A_i) = \frac{\sum_{ij} w_{ij} \left| \frac{1}{|A_i|^{p-1}} - \frac{1}{|\bar{A}_i|^{p-1}} \right|^p}{\left| \frac{1}{|A_i|^{p-1}} - \frac{1}{|\bar{A}_i|^{p-1}} \right|^p} = \sum_{ij} w_{ij} \left| \frac{1}{|A_i|^{p-1}} - \frac{1}{|\bar{A}_i|^{p-1}} \right|^{p-1} \leq \frac{\sum_{ij} w_{ij} \left| \frac{2}{\min\{|A_i|, |\bar{A}_i|\}} \right|^{p-1}}{\left| \frac{1}{|A_i|^{p-1}} - \frac{1}{|\bar{A}_i|^{p-1}} \right|^{p-1}} = 2^{p-1} \frac{\text{cut}(A_i, \bar{A}_i)}{\min\{|A_i|, |\bar{A}_i|\}} \quad (15)$$

结合齐格切的目标函数可知

$$\lim_{p \rightarrow 1} F_p(f, A_i) = \text{Ccut}(A_i, \bar{A}_i) \quad (16)$$

即 $F_p(f)$ 的解是齐格切的一种放松的近似解, 则最优的最小化目标函数可以表示为

$$\lambda_p = \arg \min_{p=1} F_p(f) \quad (17)$$

其中: λ_p 是特征向量 f 所对应的特征值。综上分析可知, p-Laplacian 的引入可以使谱聚类获得齐格图切判据的近似解, 即使某一类在聚类后所占的空间规模与该类中数据点规模的相关性尽量小, 从而获得较好的降维结果。

2.2 基于 p-Laplacian 的维数约简算法

设高维空间 $\mathbb{R}^D$ 中的数据 $X = \{x_1, x_2, \dots, x_N\}$, 令 M 是 $\mathbb{R}^D$ 中的一个嵌入流形, 在 $x_i \in \mathbb{R}^D$ 的条件约束下, 包含寻求数据点在低维空间 $\mathbb{R}^D$ 中的一组数据点 $Y = \{y_1, y_2, \dots, y_N\}$ 。利用改

进后的谱聚类来构建维数约简算法时,样本间的亲和度矩阵选用下面的方法,即在由样本集构建的图 $G(V, E)$ 中,设

$$\begin{cases} w_{ij} = \max\{\theta_i(j), \theta_j(i)\} & x_i \text{ 和 } x_j \text{ 相连} \\ w_{ij} = 0 & x_i \text{ 和 } x_j \text{ 不相连} \end{cases} \quad (18)$$

其中: $\theta_i(j) = \exp\left\{-\frac{4\|x_i - x_j\|^2}{\sigma_i^2}\right\}$, σ_i 为 x_i 到它的近邻点之间的距离。

根据以上分析,基于 p-Laplacian 的谱聚类维数约简算法流程如下所示:

输入: 数据集 $X = \{x_1, x_2, \dots, x_N\}$, 聚类数 k , 参数 $p = 1.2$ 。

输出: 样本在低维空间的投影 $Y = \{y_1, y_2, \dots, y_N\}$, 完成降维。

a) 根据式(18)构造样本亲和度矩阵 $W \in \mathbb{R}$;

b) 初始化聚类 $A_1 = V$, 聚类数 $s = 1$;

c) 重复步骤 c) ~ g);

d) 对每一个聚类 $A_i (i = 1, 2, \dots, s)$, 最小化目标函数 $F_p(f)$;

e) 利用式(7)或(8)对非规范矩阵或规范矩阵求解最优的边值,从而得到齐格切判据的近似解;

f) 利用切割判据的最小目标函数对 $A_i (i = 1, \dots, s)$ 进行分割;

g) $s \leftarrow s + 1$;

h) $s = k$, 循环结束;

i) 利用过程中求解的特征值所对应的特征向量来张成低维空间,得到样本在低维空间的投影。

3 实验仿真与结果分析

为了分析本文提出的谱聚类维数约简算法的降维性能,分别选取了人脸数据集和 UCI 数据集进行了以下实验。实验的计算机环境为:处理器为 Pentium® 3.00 GHz,内存 2 GB,硬盘 500 GB,操作系统为 Windows XP,编程语言为 MATLAB 2008b。

3.1 算法的有效性分析

影响谱聚类算法的一个重要因素是选择合适的相似度矩阵,一个理想的相似度矩阵应该是块对角矩阵,即对于一个聚类 $\{A_1, \dots, A_k\}$,当数据点 x_i, x_j 属于不同类时 $s_{ij} = 0$,则称矩阵 S 为对应于聚类 $\{A_1, \dots, A_k\}$ 的块对角矩阵。实验选择 circles (三个圆环)数据集,对矩阵按照聚类属性重新排序。图 1 显示了在原始空间计算的相似性矩阵和经 p-Laplacian 谱映射后得到的相似性矩阵。结果显示,原始空间的相似性矩阵无法体现各类数据间的区分度,而经过 p-Laplacian 谱映射后得到的相似性矩阵显示了明显的块对角模式,从而说明了该方法可增大类内相似度,同时缩小类间相似度。

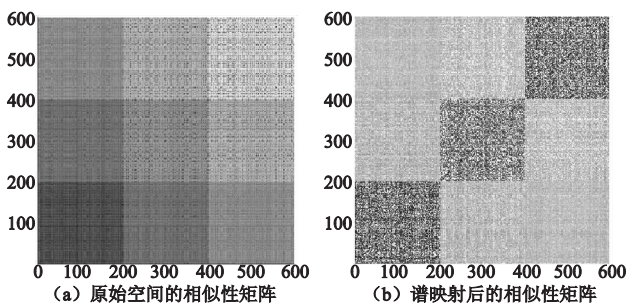


图1 三个圆环数据集上的相似性矩阵

3.2 人脸数据集

实验选用 image data 中的 CMU face image 数据集,图 2 显示了人脸数据集中的部分人脸数据。对每张人脸图片,取其灰度值作为输入特征。对 p-Laplacian 参数 p 的考虑,分别采用 $p = \{1.1, 1.2, 1.5, 1.7\}$ 四个参数值进行了多次实验,根据实验结果最终调整参数 $p = 1.2$,然后把改进后的拉普拉斯矩阵的第二特征向量和第三特征向量分别作为嵌入空间的 x 轴和 y 轴。



图2 CMU人脸数据集示例

通过 PLSC 算法计算得到了低维嵌入数据点,然后用部分具有代表性的原人脸图像代替嵌入空间的点,则可得到如图 3 所示的直观的低维嵌入结果。该图中同类图像分布得比较紧凑,不同类相互间则很分散,直观地说明了在降维过程中使用 PLSC 方法可以得到很好的低维嵌入结果。

3.3 UCI 数据集

为进一步验证本文方法的有效性,选择了具有高维特征的 UCI 数据集进行实验分析,表 1 给出了数据集的属性。对比方法包括 LDA、LLE、ISOMAP 以及 PLSC。LDA 是一种经典的线性降维方法;LLE 是一种无监督学习算法,可以最大限度地保持线性组合系数的线性子空间嵌入;ISOMAP 可以获得最大限度地保持测地线距离的线性子空间嵌入,这些算法在维数约简领域都引起了广泛的关注,具有较高的代表性,所以本文选用这三种算法作为对比算法。实验首先用四种算法对 UCI 数据集进行降维,然后用半监督学习算法识别,最后对识别正确率进行比较。识别时采取 10 倍交叉验证的方法,选取数据集中的 10% 作为训练样本,剩余作为测试样本,得出训练集对测试集的识别正确率,进行多次实验,将平均结果作为最终的识别结果。

表1 UCI数据集属性

数据集	类别数	特征维数	样本数
Ionosphere	3	80	300

表 2 和图 4 为实验结果。

表2 UCI数据集平均正确识别率比较

算法	LDA	LLE	ISOMAP	PLSC
识别率/%	81.37	86.79	87.34	90.18

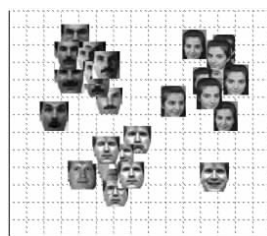


图3 人脸数据集上的低维嵌入结果

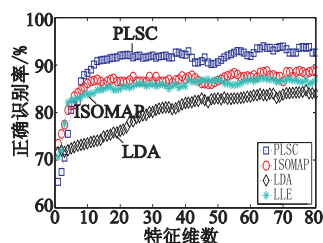


图4 UCI数据集上的降维性能比较

LDA 算法为线性方法,为了使其能够处理非线性高维数据,首先把数据非线性地映射到某个特征空间,然后在这个特征空间中进行 Fisher 线性判别,这样就隐含地实现了对原输入空间的非线性判别,从而可以使之和其他三种非线性方法具有可比性。从该实验结果中可以看出,LDA 算法的降维效果较差,ISOMAP 稍优于 LLE,但差别不大,而 PLSC 与其他算法相比效果较优。这是因为经过对拉普拉斯矩阵的非线性一般化,算法能够得到一个优化图切的近似解,并在嵌入空间中使得类内更集中,类间更分散,获得较好的降维性能,从而提高了识别的正确率。

(下转第 2131 页)

4 结束语

利用谱聚类来构造非线性维数约简算法,可以得到数据点在低维空间的嵌入,从而能很好地实现维数约简。本文提出基于 p -Laplacian 的谱聚类维数约简算法,获得了一种齐格图切的近似解,并通过对实验结果的分析表明,该算法使得投影到变换后的子空间中,类间的数据点具有更好的区分度。与其他降维算法相比,在 UCI 数据集上的识别率也有着较好的效果。由于在该方法中需要调整参数 p 的大小,给 PLSC 算法带来了一定的影响,将会在下一步的工作中对参数 p 的调整进行深入研究。

参考文献:

[1] 谭璐. 高维数据的降维理论及应用[D]. 杭州:浙江大学,2005.

- [2] Von LUXBURG U. A tutorial on spectral clustering[J]. *Statistics and Computing*,2007,17(4):395-416.
- [3] AMGHIBECH S. Bounds for the largest p -Laplacian eigenvalue for graphs[J]. *Discrete Mathematics*,2006,306(21):2762-2771.
- [4] HEIN M, AUDIBERT J Y, Von LUXBURG U. Graph Laplacian and their convergence on random neighborhood graphs[J]. *Journal of Machine Learning Research*,2007,8(12):1325-1368.
- [5] FLECKINGER J, HARRELL II E M, De THELIN F. On the fundamental eigenvalue ratio of the p -Laplacian[J]. *Bulletin des Sciences Mathematiques*,2007,131(7):613-619.
- [6] WANG Xiang, DAVIDSON I. Active spectral clustering[C]//Proc of IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society,2010:561-568.
- [7] BINDING P A, RYNN E B P. Variational and non-variational eigenvalues of the p -Laplacian[J]. *Differential Equations*,2008,244(1):24-39.