第 37 卷 第 9 期 计 算 机 工 程 2011 年 5 月
Vol.37 No.9 Computer Engineering May 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)09—0216—02 文献标识码：A 中图分类号：TP18

基于拉普拉斯特征映射的仿射传播聚类

张 亮，杜子平，张 俊，李 杨

(天津科技大学经济与管理学院，天津 300222)

摘 要：仿射传播方法难以处理具有流形结构的数据集。为此，提出一种基于拉普拉斯特征映射的仿射传播聚类算法(APPLE)，在标准仿射传播的基础上增强流形学习的能力。使用测地距离计算数据点间相似度，采用拉普拉斯特征映射对数据集进行降维及特征提取。对图像聚类应用的实验结果证明了 APPLE 的聚类效果优于标准仿射传播方法。

关键词：拉普拉斯特征映射；仿射传播；Dijkstra 算法；归一化互信息

Affinity Propagation Clustering Based on Laplacian Eigenmaps

ZHANG Liang, DU Zi-ping, ZHANG Jun, LI Yang

(School of Economics and Management, Tianjin University of Science and Technology, Tianjin 300222, China)

【Abstract】Affinity propagation is often limited by its inability to cluster datasets with inherent manifold structures. A novel clustering method, namely Affinity Propagation with Laplacian Eigenmaps(APPLE), is proposed to address this problem. It enhances the standard affinity propagation with manifold learning capacity. Geodesic distance is used to compute affinity between data points. Laplacian eigenmaps are applied to reduce the dimensionality and to extract features. Experimental results show APPLE outperforms standard affinity propagation in application of image clustering.

【Key words】Laplacian eigenmaps; Affinity Propagation(AP); Dijkstra algorithm; Normalized Mutual Information(NMI)

DOI: 10.3969/j.issn.1000-3428.2011.09.075

1 概述

聚类是机器学习的主要研究任务之一。它的目的是将数据点按照相关性分成若干组，从中探寻数据集内在的结构信息。聚类方法被广泛应用于模式识别、数据挖掘、图像分析、计算生物学等多个领域。

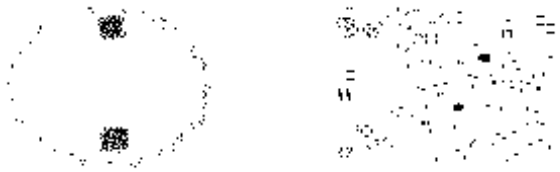
针对不同的应用领域以及具有不同特点的数据集，近期聚类方法的热点研究方向包括：基于模型的聚类，核聚类，图聚类和谱聚类，相似度学习等。尽管多种聚类理论和方法得到了快速发展，但大多数聚类算法仍然面临着可扩展性差、需要人工预先确定类别数、对数据集的内在流形特征缺乏有效的表示和提取能力、难以处理高维数据等许多具体问题。

文献[1]提出一种基于消息扩散机制的仿射传播(Affinity Propagation, AP)聚类方法，以其能够迅速、有效地处理大规模聚类问题的优秀特性受到研究者的广泛关注[2]。大部分 K-中心型算法的聚类效果对初始中心点即“代表点”(examplar)的选择敏感，而 AP 算法很好地避免了这一问题。通过在数据点间构造递归的消息传播机制，AP 能够以相当高的计算效率自动收敛到代表点。然而，在很多实际聚类问题中，AP 算法仍有一些值得改进的方面：(1)数据集中往往包含线性不可分的结构，标准的 AP 算法对这类问题难以得到准确的聚类。(2)对某些高维数据集，直接计算得到的数据点间的相似度并不能准确反映数据集的内在流形特征，可能会降低 AP 算法的性能。往往需要预先采取某种降维手段将数据点映射到低维特征空间上，以提高聚类方法的准确性。

本文对标准 AP 算法的不足进行改进，提出一种新的基于拉普拉斯特征映射的仿射传播聚类(Affinity ProPagation with Laplacian Eigenmaps, APPLE)方法。引入一种测地距离映射方法对原始数据集进行降维，在映射空间上应用 AP 算法进行聚类。

2 仿射传播聚类

图 1 显示了基于欧氏距离的 AP 算法聚类结果。



(a)线性不可分的数据 (b)密度不均匀的数据

图 1 基于欧氏距离的 AP 算法聚类结果

记数据集为 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\mathrm{L}\ ,\boldsymbol{x}_N\}$，$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间的距离为 $d(\boldsymbol{x}_i,\boldsymbol{x}_j)$。对于 K-中心型的距离问题，目的是寻找 K 个代表点，记为 $X_e=\{x_{e1},x_{e2},\mathrm{L}\ ,x_{eK}\}$，使得其他点与代表点间的距离最小化：

$$D(X,X_e)=\sum_{i=1}^{N}d(\boldsymbol{x}_i,e(\boldsymbol{x}_i)) \tag{1}$$

其中，$e(\boldsymbol{x}_i)$ 是 $\boldsymbol{x}_i$ 的代表点。每个代表点是一个实际的数据点，是其所在聚类的中心。AP 聚类的基本思想是寻找这些代表

基金项目：国家自然科学基金资助项目(70671074)；天津市科技发展战略研究计划基金资助项目(10ZLZLZF04900)

作者简介：张 亮(1979－)，男，讲师、博士，主研方向：机器学习，人工智能；杜子平，教授、博士、博士生导师；张 俊、李 杨，讲师、博士

收稿日期：2010-10-18 **E-mail：**zhangliang@tust.edu.cn

点，并将其他数据按照代表点划分到相应的类中。

在 AP 方法中，数据点间交换 2 类消息：代表度(responsebility) $r(i,k)$和有效度(availability) $a(i,k)$。前者由数据点 $\boldsymbol{x}_i$ 发送到候选代表点 $\boldsymbol{x}_k$，反映 $\boldsymbol{x}_k$ 作为 $\boldsymbol{x}_i$ 代表点的适合程度；后者由由候选代表点 $\mathbf{x}_k$ 发送到数据点 $\boldsymbol{x}_i$，反映 $\boldsymbol{x}_i$ 选择 $\boldsymbol{x}_k$ 作为其代表点的适合程度。初始化 $a(i,k)=0$，$r(i,k)=0$，然后迭代地计算：

$$r(i,k)=s(i,k)-\max_{k',k'\neq k}\{a(i,k')+s(i,k')\} \tag{2}$$

$$r(k,k)=s(k,k)-\max_{k',k'\neq k}\{s(k,k')\} \tag{3}$$

$$a(i,k)=\min\{0,r(k,k)+\sum_{i',i'\neq i,k}\max\{0,r(i',k)\}\} \tag{4}$$

$$a(k,k)=\sum_{i',i'\neq i}\max\{0,r(i',k)\} \tag{5}$$

其中，$s(i,k)$反映了数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_k$ 的相似度，通常取欧氏距离的负平方值 $s(i,k)=-\|\boldsymbol{x}_i-\boldsymbol{x}_k\|^2$。而 $s(i,i)$是 $\boldsymbol{x}_i$ 的“偏好值”，一个数据点的偏好值越高，其被选为代表点的可能性越大。当迭代过程收敛时，对每个数据点 $\boldsymbol{x}_i$，其代表点 $e(\boldsymbol{x}_i)$ 的下标为：

$$\arg\max_k\{a(i,k)+r(i,k),k=1,2,\mathrm{L}\ ,N\} \tag{6}$$

AP 算法可以自动寻找聚类及其代表点，在处理类数较多的数据集时速度较快。然而，如图 1 所示，标准 AP 方法难以准确处理具有内在流形结构的数据。这限制了 AP 方法在图像检索、语音识别等很多流形结构特征突出的实际问题中的应用。

3　基于测地距离的拉普拉斯特征映射

AP 算法通过数据图中节点之间的消息传播，由局部信息的扩散来实现全局学习。而流形学习的目标是在保持局部结构信息的同时实现全局学习器的构建。两者的学习方式非常接近。本文希望借助 AP 算法局部相关信息传播机制实现对具有流形结构的数据进行聚类。

文献[3]的研究表明，当数据集具有某种内在流形结构时，测地距离与欧氏距离相比能够得到更好的聚类效果。文献[4]验证了在谱聚类和拉普拉斯映射算法中，采用测地距离度量与欧氏距离相比有突出的优势。因此，考虑采用测地距离取代欧氏距离作为数据点间相似度的度量形式。在一个图中，两点之间的测地距离由连接它们的最短路径表示。图 2 显示了采用测地距离计算相似度，AP 算法可以得到更准确的聚类效果。

(a)线性不可分的数据　　(b)密度不均匀的数据

图 2　基于测地距离的 AP 算法聚类结果

图 2 中任意两点间测地距离的计算可采用 Floyd 算法或 Dijkstra 算法。同时，结合流形学习领域常用的拉普拉斯特征映射[5]方法，可将数据集 X=$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\mathrm{L}\ ,\boldsymbol{x}_N\}$ 由高维原空间映射到低维特征空间上。将这种方法称为基于测地距离的拉普拉斯特征映射，具体算法为：

(1)由 X=建立一个最近邻图 G。在所有 N 个数据点上定义图，其中若两点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间的距离小于 ε 或互为 k 个最近邻点之一，则在它们之间连接一条边，边长度定义为欧氏距离 $d_E(\boldsymbol{x}_i,\boldsymbol{x}_j)$。

(2)计算最短路径矩阵 $\boldsymbol{D}_G$ 和相似度矩阵 $\boldsymbol{W}=\{w_{ij}\}$。如果 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 直接相连，初始化 $d_G(\boldsymbol{x}_i,\boldsymbol{x}_j)=d_E(\boldsymbol{x}_i,\boldsymbol{x}_j)$，否则 $d_G(\boldsymbol{x}_i,\boldsymbol{x}_j)=\infty$。然后将每个 $d_G(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 替换为 $\min(d_G(\boldsymbol{x}_i,\boldsymbol{x}_j),d_G(\boldsymbol{x}_i,\boldsymbol{x}_k)+d_G(\boldsymbol{x}_k,\boldsymbol{x}_j))$，$k$=1,2,⋯,$N$。矩阵 $\boldsymbol{D}_G=(d_G(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1}^N$ 存储图中任意两点间的测地距离。设置 $w_{ij}=\exp(-d_G^2(\boldsymbol{x}_i,\boldsymbol{x}_j)/\sigma^2)$，其中，$\sigma$ 是尺度参数，反映了测地距离高斯分布函数的宽度。文献[6]给出一种确定 σ 的方法：令 $\sigma^*\in\{0.001,0.01,0.1,1,10,100,1\ 000\}$，$\sigma=1/\sqrt{2\sigma^*}$。在第 4 节实验中，将对每个数据集分别采用 5 折交叉检验来确定 σ 的值。

(3)计算数据点在低维空间的映射集 Z $=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\mathrm{L}\ ,\boldsymbol{z}_N\}$。定义 $\boldsymbol{D}=\{\sum_j w_{ij}\}$，图的拉普拉斯矩阵为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，计算方程 $\boldsymbol{Lf}=\lambda\boldsymbol{Df}$ 的特征值和特征向量。将特征值 λ 按从小到大的顺序排序为 $\lambda=\{\lambda_1,\lambda_2,\mathrm{L}\ ,\lambda_N\}$，各特征值对应的特征向量为 $\boldsymbol{f}=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\mathrm{L}\ ,\boldsymbol{f}_N\}$。取第 2 到 m+1 个特征向量作为原空间到特征空间的映射函数，则有 $\boldsymbol{x}_i\rightarrow\boldsymbol{z}_i=(\boldsymbol{f}_2(i),\boldsymbol{f}_3(i),\mathrm{L}\ ,\boldsymbol{f}_{m+1}(i))$。其中，$\boldsymbol{z}_i$ 即为数据点 $\boldsymbol{x}_i$ 对应在低维特征空间 Z 上的坐标。

(4)对低维特征空间上的数据集 Z 应用 AP 算法，得到聚类结果。

4　实验与分析

本文方法在 Matlab 软件中实现。在 2 个图像数据集 COIL-20 和 COREL-1 000 上进行聚类效果的测试。COIL-20 包括 20 类，每类包含 72 张关于同一 3D 物体不同角度的图像。对每张图片，调整大小为 32×32，得到 1 024 维数据特征。

COREL-1 000 包括 10 类，每类由 100 张同类图像组成。对每张图片，取 32 维($H\times S$:8×4)色彩直方图和 16 维共现纹理，共 48 维作为特征。

评价指标采用归一化互信息(Normalized Mutual Information, NMI)，计算方法是：

$$N_{\mathrm{NMI}}=\frac{MI(C,C')}{\max(H(C),H(C'))} \tag{7}$$

其中，C 是实际类；C’是由聚类算法产生的类；$MI(C,C')$是互信息度量；$H(C)$和 $H(C')$分别是 C 和 C’的熵。N_{NMI} 是一个 0,1 间的值，它的值越大说明聚类效果越好。

在 2 个图像数据集上，分别应用 3 种基于图的聚类算法：mincut[7]、AP[1]以及本文提出的 APPLE 方法。对每种方法分别采用欧氏距离度量(Euclidean Distances, ED)和测地距离度量(Geodesic Distances, GD)进行实验。APPLE 方法中对原始数据集进行降维。

在 2 个数据集上，首先运行第 3 节中基于测地距离的拉普拉斯特征映射算法，分别将原始的 1024 维和 48 维空间降为 16 维特征空间，然后进行 AP 聚类。AP 和 APPLE 的最大迭代次数设置为 500 次。

表 1 显示了这 3 种方法的聚类结果。

表 1　3 种方法的聚类结果

数据集	Mincut		AP		APPLE	
	ED	GD	ED	GD	ED	GD
COIL-20	0.785	0.791	0.779	0.821	0.903	0.906
COREL-1000	0.672	0.714	0.631	0.710	0.727	0.735

从表 1 可以得到以下结论：

(下转第 220 页)