DOI: 10.11999/JEIT151132

基于 Kullback-Leiber 距离的迁移仿射聚类算法

毕安琪* 王士同

(江南大学数字媒体学院 无锡 214122)

摘 要:针对迁移聚类问题,该文提出一种新的基于 Kullback-Leiber 距离的迁移仿射聚类算法(TAP_KL)。该算法从概率角度重新解释 AP 算法的目标函数,并借助于信息论中最常见的一种距离度量,即 Kullback-Leiber 距离,测量源域与目标域代表点的相似性。另外,通过详细分析 TAP_KL 算法与 AP 算法的目标函数,得出一个重要结论,即可以将源域与目标域的相似性嵌入到目标域数据集相似性矩阵的计算中,从而直接利用 AP 算法的优化算法优化 TAP_KL 算法的目标函数,解决基于代表点的迁移聚类问题。最后,通过基于 4 个数据集的仿真实验,进一步验证了 TAP_KL 算法在解决迁移聚类问题时的有效性。
 关键词:仿射聚类算法;迁移学习;人脸数据集;概率框架;KL 距离
 中图分类号: TP391.4
 文章编号: 1009-5896(2016)08-2076-09

Transfer Affinity Propagation Clustering Algorithm Based on Kullback-Leiber Distance

BI Anqi WANG Shitong

(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract: For solving the clustering problem of transfer learning, a new algorithm called Transfer Affinity Propagation clustering algorithm is proposed based on Kullback-Leiber distance (TAP_KL). Based on the probabilistic framework, a new interpretation of the objective function of Affinity Propagation (AP) clustering algorithm is proposed. By leveraging Kullback-Leiber distance which is usually used in information theory, TAP_KL measures the similarity relationship between source data and target data. Moreover, TAP_KL algorithm can embed the similarity relationship to the calculation of similarity matrix of target data. Thus, the optimization framework of AP can be directly used to optimize the new target function of TAP_KL. In this case, TAP_KL builds a simple algorithm framework to solve the transfer clustering problem, in which the algorithm just needs to modify the similarity matrix to solve the transfer clustering problem. The experimental results based on both 4 datasets show the effectiveness of the proposed algorithm TAP_KL.

Key words: Affinity Propagation(AP) clustering algorithm; Transfer learning; Face datasets; Probabilistic framework; Kullback-Leiber distance(KL)

1 引言

近年来,国内外研究学者从不同角度对迁移学 习的研究已经取得了众多重要研究成果^[1-8],包括 迁移 SVM(Support Vector Machine)算法^[1,2]、迁移 Adaboost 算法^[3-4],以及基于流形结构的 MMDE (Maximum Mean Discrepancy Embedding)算 法^[5-6]。然而聚类算法作为机器学习和模式识别领域的一个重要研究方向,现阶段对于迁移聚类算法的研究并不充分,取得的成果也不多^[8]。聚类算法的目标是将相似的数据聚集为一个数据簇,并使差异较大的数据分别属于不同的数据簇。目前广泛使用的聚类算法,包括 K-均值算法^[0,10]、谱聚类算法^[11,12]、仿射聚类(Affinity, Propagation, AP)^[13-19]以及模糊聚类方法^[20]都是在数据量足够充分的前提下,才能保证算法得到可靠的、有效的聚类结果。因此,这些算法都不适用于迁移学习的场景中,本文就是针对迁移聚类问题进行研究探讨。

聚类算法的一个重要研究方向就是从已经存在 的样本点中选择算法所得的数据簇类中心,这类算 法统称为基于代表点的聚类算法,其中最具代表性

收稿日期: 2015-10-10; 改回日期: 2016-04-17; 网络出版: 2016-06-03 *通信作者: 毕安琪 angela.sue.bi@gmail.com

基金项目:国家自然科学基金(61170122, 61272210),江苏省 2014 年度普通高校研究生科研创新计划项目(KYLX_1124),山东省高 等学校科技计划项目(J14LN05)

Foundation Items: The National Natural Science Foundation of China (61170122, 71272210), Jiangsu Graduate Student Innovation Projects (KYLX_1124), The Science and Technology Program Shandong Provinceial Higher Education (J14LN05)

的算法包括 AP 算法^[13], EEM 算法(Enhanced α-Expansion Move)^[18,19]等。研究指出,基于代表点的聚类算法的目标函数均可以看作是马尔科夫随机场(Markov Random Field, MRF)的能量函数,相应地,AP 算法与 EEM 算法本质上是在优化相同的目标函数,其中 AP 算法使用的优化算法是基于样本点之间的信息传递,而 EEM 算法则基于 Graph-Cuts。另一方面,基于代表点的聚类算法的一个重要优势在于算法可以根据数据集自动完成聚类,而不要求预设数据簇的总数。

对于现有的迁移学习算法,一个重要的问题是 如何基于源域与目标域数据的相似性,更好地借助 于源域数据的研究成果完成对目标域数据的研究。 在基于代表点的聚类算法中,进一步可以认为源域 与目标域的相似性表现为源域与目标域代表点集合 的相似性。具体地,本文首先从概率角度重新解释 AP 算法的目标函数,通过定义样本点和代表点的概 率关系,以及代表点集合的先验概率,为度量源域 与目标域代表点集合的相似性提供前提条件;其次, 借助于信息论中最常见的一种距离度量,即 Kullback-Leiber 距离(KL 距离),在概率框架下测量 源域与目标域代表点的相似性,提出了 TAP KL 算法的目标函数;最后,通过详细分析 TAP KL 算法与 AP 算法的目标函数,得出一个重要结论, 即可以将源域与目标域的相似性嵌入到目标域数据 集相似性矩阵的计算中,从而直接借助 AP 算法的 优化策略解决新的目标函数。

2 AP 算法

2007 年, 文献[13]中指出, 在聚类算法中, 若 取得的类中心点是从已存在的样本点中选择的, 则 称这类聚类算法为基于代表点聚类算法, 并称这些 类中心点为代表点。同时, 文献[13]提出一种典型的 基于代表点聚类算法, 即 AP 聚类算法, 其目标函 数可以表示为

$$\begin{split} \min_{\boldsymbol{c}} \sum_{p=1}^{N} d\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{c_{p}}\right) + \sum_{p=1}^{N} \delta_{p}(\boldsymbol{c}), \\ \delta_{p}(\boldsymbol{c}) = \begin{cases} \infty, & \boldsymbol{c}_{p} \neq p , \exists \boldsymbol{x}_{q} : \boldsymbol{c}_{q} = p \\ 0, & \nexists \dot{\boldsymbol{c}} \end{cases} \end{split}$$
(1)

其中, $X = \{x_1, x_2, \dots, x_p, \dots, x_N\} \in \mathbb{R}^{N \times D}$ 是一个包含 N个 D 维样本的数据集, $c = \{c_1, c_2, \dots, c_p, \dots, c_N\}$ 表 示待求的数据集样本点下标集合。目标函数中第 2 项 $\delta_p(c)$ 保证了所得代表点集合的有效性, 如果某代 表点 $c_p \in c$ 所选择的代表点不是 p ,此时代表点集 合是无效的,即若当前代表点集合 c 有效,则 $\forall c_p$ $\in c, c_{c_p} = c_p$ 。另外, $d(x_p, x_{c_p})$ 为样本点与代表点 之间的欧氏距离,在 AP 算法中使用欧氏距离度量 样本间的相似性 $s(\mathbf{x}_{p}, \mathbf{x}_{c_{p}})$,即认为样本间欧氏距离 越小,其相似度越高。

将上述目标函数的优化过程可以看作 MRF 的 能量函数最小化过程,事实上,所有的基于代表点 聚类算法的目标函数优化问题都可以看作是 MRF 的能量函数寻优问题。AP 算法使用基于信息传递的 LBP(Loopy Belief Propagation)优化算法来优化式 (1)。在具体优化过程中,算法首先定义 2 个矩阵 R = (r(i,k))和A = (a(i,k))分别存储样本点传递给 代表点的信息和代表点传递给样本点的信息,具体 定义为

$$r(i,k) \leftarrow s(i,k) - \max_{j' \neq k} \{a(i,j') + s(i,j')\},\$$
$$a(i,k) \leftarrow \left\{ \min_{i \neq k} \left\{ 0, r(k,k) + \sum_{j \neq \{i,k\}} \{\max(0, r(j,k))\} \right\}, \quad i \neq k$$
$$\sum_{j \neq \{k\}} \{\max(0, r(j,k))\}, \quad i = k$$
(2)

其中, S = (s(i,j))为 N个样本之间的相似性矩阵, 根据 AP 的目标函数式(1),式(2)中 $s(i,j) = -d(\mathbf{x}_i,\mathbf{x}_j)$,设 $s(i,i) = \alpha$ 是一个常数,称为偏向参数 (preference),其值将由用户定义,根据文献[13], α 的值将影响算法产生的数据簇总数。一般来说, α 值 越大,算法产生的类数越少,反之亦然。另一方面, 在第 m 次迭代过程中, $\mathbf{R}_m \cap \mathbf{A}_m$ 需要进行加权更 新,更新公式为

$$\begin{aligned} \mathbf{R}_{m} &= (1 - l_{d})\mathbf{R}_{m} + l_{d}\mathbf{R}_{m-1} \\ \mathbf{A}_{m} &= (1 - l_{d})\mathbf{A}_{m} + l_{d}\mathbf{A}_{m-1} \end{aligned}$$
(3)

其中, $l_d \in [0,1]$ 称为阻尼系数,一般设置 $l_d = 0.5$ 。 阻尼系数对于增强算法的收敛性起着重要作用,当 算法发生震荡时,增大 l_d 可以帮助算法收敛。

AP 算法不需要提前预设聚类总数,算法能够根据数据集的相似性矩阵,自动计算出合适的聚类总数,从而获取有效的聚类结果;另一方面,实验证明 AP 算法所得到的聚类性能相当有效与稳定。基于这两个优势,近年来 AP 算法得到了国内外研究者的广泛关注,并已经取得了若干重要研究成果,其中包括半监督 AP 算法^[14],递增式 AP 算法^[16,17]等。为了解决迁移聚类问题,本文在保留原始 AP 算法以上两个优势的基础上,提出了一种改进的 AP 算法,即基于 Kullback-Leiber 距离的迁移仿射聚类算法。

基于 Kullback-Leiber 距离的迁移仿射聚 类算法

迁移聚类问题涉及到两个数据集,即源域数据

集与目标域数据集。因此,如何准确地度量目标域 与源域数据集之间的相似性,以及目标域数据集样 本间的相似性,是亟待解决的问题。另一方面,概 率框架能够准确地体现数据的分布特征,而在信息 论中已知若干种距离可用来度量两个概率分布的相 似性。

因此,本文首先利用概率的信息表征特征,引入概率框架重新解释 AP 算法目标函数的合理性及 有效性,其次,在新的目标函数的基础上,利用 Kullback-Leiber 距离度量源域数据集与目标域数据 集的相似性,进而发现可以将源域与目标域的相似 性嵌入到目标域数据集的相似性矩阵的计算中,并 借助 AP 算法的优化算法解决迁移仿射聚类问题。

3.1 AP 算法的概率框架下解释

在信息论中,概率能够更好地体现数据的分布 特征。而在迁移学习中,较准确的表示源域和目标 域的数据分布是解决其他问题的基础和前提。因此, 本节首先引入相关的概率定义,然后表明在高斯概 率假设下,可诱导出等价的 AP 算法的目标函数。 换句话说,通过引入概率框架,我们可以重新解释 AP 聚类算法的目标函数,进而给出 AP 算法基于概 率框架的目标函数。该概率框架为之后解决迁移学 习中的聚类问题提供了可靠的基础。

令 E 表示代表点的下标集合,基于样本点与代表点间的相似度,定义样本点 x_p 选择 $E(x_p)$ 作为代表点的概率为

$$p\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-d\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right)/2\sigma^{2}\right) (4)$$

其次,若当前代表点集合中存在某代表点选择 除自己以外的其他代表点,则当前代表点集合无效, 即 $\forall i \in E, E(x_i) = i$ 。与 AP 算法的目标函数类似, 在概率框架下也可以通过定义 $\theta_{m,n}(E(x_m), E(x_n))$ 来 避免这类无效的代表点集合的产生。因此,代表点 集合的先验概率为

$$p(\boldsymbol{E}) = \frac{1}{\sigma\sqrt{2\pi}}$$

$$\cdot \exp\left(-\sum_{m=1}^{N}\sum_{n>1}^{N}\theta_{m,n}\left(\boldsymbol{E}(\boldsymbol{x}_{m}), \boldsymbol{E}(\boldsymbol{x}_{n})\right)/(2\sigma^{2})\right)(5)$$

$$\theta_{m,n}(\boldsymbol{E}(\boldsymbol{x}_{m}), \boldsymbol{E}(\boldsymbol{x}_{n})) = \begin{cases} \infty, & \boldsymbol{E}(\boldsymbol{x}_{m}) = n, \boldsymbol{E}(\boldsymbol{x}_{n}) \neq n, \\ & \boldsymbol{E}(\boldsymbol{x}_{n}) = m, \boldsymbol{E}(\boldsymbol{x}_{m}) \neq m \\ 0, & \boldsymbol{\Xi} \succeq \end{cases}$$

由于聚类过程要求算法找到一个有效的代表点 集合,并使以上两项概率值最大。因此,从概率角 度重新解释 AP 算法,得到的新的目标函数为

(6)

$$\max_{\boldsymbol{E}} Q(\boldsymbol{E}) = \sum_{p=1}^{N} \ln p\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right) + \ln p(\boldsymbol{E}) \qquad (7)$$

其中,N表示数据集中样本点的个数,**E**表示每个 样本点所选择的代表点下标集合。进一步简化目标 函数式(7),并忽略常数项的影响,可以认为式(7) 与 AP 算法的目标函数式(1)是等价的。

因此,本节通过从概率角度重新考虑基于代表 点的 AP 聚类算法,推导出了同样的 AP 算法的目 标函数,这将为有效利用 AP 算法解决迁移聚类问 题提供了前提条件;也就是在解决迁移聚类问题的 时候,可以进一步利用概率来度量源域与目标域样 本的关系。

3.2 基于 Kullback-Leiber 距离的迁移仿射聚类算法 — TAP KL

在试图利用 AP 聚类框架解决迁移聚类问题的 时候,选择一种可以测量源域与目标域数据分布的 距离公式具有重要作用。尽管我们可以采用其他的 方法(如卡方检验(Chi-Square), Hausdorff 距离)来 研究迁移聚类。但这里我们选择 KL 距离,其原因 是:(1)KL 距离是信息论中最常见的一种距离度量 方法。(2)基于 KL 距离,我们发现所得到的目标函 数可以直接使用 AP 算法的优化算法,而不需要重 新构建新的优化算法。这也是本文的贡献之一。KL 距离是信息论中最常见的一种距离度量,这种距离 从统计学角度测量两个概率分布之间的相似性。假 设存在两个概率分别为 P和 Q,则概率分布 P 到概 率分布 Q 的 KL 距离定义为

$$D_{\mathrm{KL}}\left(P \| Q\right) = \sum_{\boldsymbol{x} \in \boldsymbol{X}} P(\boldsymbol{x}) \ln \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})}$$
(8)

其中, $X = \{x_1, x_2, \dots, x_n\}$ 表示样本点集合。需要注意的是, KL 距离不是对称距离, 即 $D_{\text{KL}}(P || Q) \neq D_{\text{KL}}(Q || P)$ 。

基于代表点的迁移聚类算法要求目标域产生的 代表点集合与源域的代表点集合尽可能相似。具体 地,首先目标域中的样本点*x*_p从源域代表点集合中 选择最合适的代表点*L*(*x*_n)

$$\boldsymbol{L}(\boldsymbol{x}_{p}) = \underset{\boldsymbol{x}_{l} \in \boldsymbol{L}_{s}}{\operatorname{argmax}} p\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{l}\right) \tag{9}$$

其中, $L_s = \{x_{si} | i \in E_s\}$ 是源域的代表点集合, $X_s = \{x_{s1}, x_{s2}, \dots\}$ 是源域数据集, E_s 代表由源域数 据产生的代表点在源域数据集中的下标集合。

其次,要找出目标域样本的代表点 $E(x_p)$,并 计算 $E(x_p)$ 与代表点的概率关系 $p(x_p, E(x_p))$ 。最后, 算法要求 $p(x_p, E(x_p))$ 与 $p(x_p, L(x_p))$ 的 KL 距离尽 可能小。综上所述,结合基于概率框架的 AP 算法 的目标函数式(7),定义基于 KL 距离的迁移仿射聚 类 (Transfer Affinity Propagation based on Kullback-Leiber distance, TAP_KL)算法的目标函数如式(10):

$$\max_{\boldsymbol{E}} Q(\boldsymbol{E}) = \ln p(\boldsymbol{E}) + \sum_{p=1}^{N} \ln p\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right) \\ - \lambda \sum_{p=1}^{N} p\left(\boldsymbol{x}_{p}, \boldsymbol{L}(\boldsymbol{x}_{p})\right) \\ \cdot \left[\ln p(\boldsymbol{x}_{p}, \boldsymbol{L}(\boldsymbol{x}_{p})) - \ln p\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right)\right]$$
(10)

其中, $x_p \in X$ 是目标域数据, $X = \{x_1, x_2, \dots, x_p, \dots, x_N\} \in \mathbb{R}^{N \times D}$ 是一个包含 $N \land D$ 维样本的目标域数据集, E 是每个样本点的代表点下标,即样本点 x_i 的代表点下标为 $E(x_i)$,其选择的代表点则为 $x_{E(x_i)}$, $p(x_p, x_{E(x_p)}), p(E)$ 和 $L(x_p)$ 分别如式(4),式(5),式(9)所示, λ 为正则化参数,考虑到 $p(x_p, L(x_p)) \in [0,1]$,即第2,3项属于同一个量纲,因此 λ 的值并不大。

进一步简化式(10),并忽略对优化过程无影响 的常数项,得到目标函数为

$$\max_{\boldsymbol{E}} Q(\boldsymbol{E}) = \ln p(\boldsymbol{E}) + \sum_{p=1}^{N} \left[1 + \lambda p\left(\boldsymbol{x}_{p}, \boldsymbol{L}\left(\boldsymbol{x}_{p}\right)\right) \right]$$
$$\cdot \ln p\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right)$$
(11)

上述目标函数式(11)与 AP 算法的目标函数具有高度的相似性,因此可以借鉴 AP 算法的优化算法来优化式(11)。

3.3 TAP KL 与 AP 算法

将 $p(\boldsymbol{E})$ 与 $p(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})})$ 的定义代入 TAP_KL 的 目标函数式(11)中,可得式(11)等价于 max $Q(\boldsymbol{E}) = -\sum_{p=1}^{N} \left[1 + \lambda p\left(\boldsymbol{x}_{p}, \boldsymbol{L}\left(\boldsymbol{x}_{p}\right)\right) \right] d\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right)$ $-\sum_{m=1}^{N} \sum_{n=1}^{N} \theta_{m,n}\left(\boldsymbol{E}(\boldsymbol{x}_{m}), \boldsymbol{E}(\boldsymbol{x}_{n})\right)$ (12)

另一方面,从样本间相似性角度来说, AP 算法的目标函数可以表示为

$$\max_{\boldsymbol{E}} \sum_{p=1}^{N} s\left(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}\right) - \sum_{m=1}^{N} \sum_{n=1}^{N} \theta_{m,n}\left(\boldsymbol{E}\left(\boldsymbol{x}_{m}\right), \boldsymbol{E}\left(\boldsymbol{x}_{n}\right)\right) (13)$$

其中, $s(\boldsymbol{x}_p, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_p)}) = -d(\boldsymbol{x}_p, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_p)}), \theta_{m,n}(\boldsymbol{E}(\boldsymbol{x}_m), \boldsymbol{E}(\boldsymbol{x}_n))$

如式(6),用以保证所求代表点集合的有效性。另一方面,由于在优化目标函数式(12)的过程中, λ 和 $p(\boldsymbol{x}_p, \boldsymbol{L}(\boldsymbol{x}_p))$ 均不变化,可认为这两项是常数项。因此,当源域的代表点集合 \boldsymbol{L}_s 已知,若定义另一种相似性度量方法为

$$s(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})}) = -[1 + \lambda p(\boldsymbol{x}_{p}, \boldsymbol{L}(\boldsymbol{x}_{p}))]d(\boldsymbol{x}_{p}, \boldsymbol{x}_{\boldsymbol{E}(\boldsymbol{x}_{p})})$$
(14)
式(13)则等价于 TAP_KL 的目标函数式(12)。

综上所述,通过使用不同的相似性度量手段, 式(13)分别扩展为 AP 算法的目标函数式(1)和式 (12)。在 AP 的优化过程中,通过定义 2 个矩阵传 递样本间的信息,2 个矩阵 $A \ \pi R$ 的定义如式(2)所 示,其迭代公式如(4), $A \ \pi R$ 均与数据集的相似性 矩阵 S 有关。因此,在优化 TAP_KL 算法的目标 函数式(14)时,只需将新的相似性矩阵 S 代入矩阵 $A \ \pi R$ 的计算中,而其他设置不变。由于新的相似 性矩阵定义中嵌入了度量源域与目标域相似性的一 项 $p(\mathbf{x}_p, \mathbf{L}(\mathbf{x}_p))$,此时优化算法得到的结果既考虑了 目标域样本间的相似性,也考虑了源域与目标域之 间的相似性。TAP_KL 算法的具体聚类步骤如表 1 所示。

表1 TAP_KL 算法

TAP_KL 算法

- **输入:** 源域数据集的代表点集合 L_s , 目标域数据集 $X = \{x_1, x_2, ..., x_p, ..., x_N\}$, 偏向参数 α , 正则化参数 λ , 阻尼系数 l_d , 最大迭代次数 M
- 输出:目标域数据集的代表点下标集合 E
- (1) for $\boldsymbol{x}_p \in \boldsymbol{X}$
- (2) 根据式(9)从源域代表点集合 L_s 中为 x_p 选择最合适的代表 点 $L(x_p)$;
- (3) 根据式(4)计算 $p(\boldsymbol{x}_p, \boldsymbol{L}(\boldsymbol{x}_p))$;
- (4) for $\boldsymbol{x}_q \in \boldsymbol{X}$
- (5) $d(\boldsymbol{x}_p, \boldsymbol{x}_p) = \alpha;$
- (6) for $\boldsymbol{x}_q \in \boldsymbol{X} \{\boldsymbol{x}_p\}$

(7) 计算
$$\boldsymbol{x}_p \models \boldsymbol{x}_q$$
的欧氏距离, $d(\boldsymbol{x}_p, \boldsymbol{x}_q) =$
 $\|\boldsymbol{x}_p - \boldsymbol{x}_q\|^2$;

- (8) end for
- (9) end for
- (10) 根据式(14),构建目标域数据集的相似度矩阵S;
- (11) t=1;
- (12) 根据式(2)以及迭代公式(3), 计算当前矩阵 R 和 A;
- (13) 获取此时代表点集合;
- (14) t = t + 1;
- (15) 算法收敛后,输出目标域数据集的代表点下标集合 E

TAP_KL 算法在解决迁移聚类问题的时候,(1) 继承了 AP 算法的优势,即不需要预设所得数据簇 总数;(2)由于将源域与目标域的相似性嵌入到目标 域相似性矩阵的计算中,TAP_KL 算法不需要重新 构建新的优化算法,而是直接利用 AP 算法的优化 算法解决迁移聚类问题,算法的时间和空间复杂度 均不会增加。因此,TAP_KL 算法在保持了与 AP 算法一致的时间与空间复杂度时,有效地解决迁移 聚类问题。

4 仿真实验

本文通过若干仿真实验,进一步验证 TAP_KL 算法的有效性,为构建合理的迁移数据集,实验中 选取了 4 个真实数据集,并将其与 AP 算法、TSC (Transfer Spectral Clustering)算法^[8]比较。

4.1 数据集与评价标准

本文主要使用两个评价指标来测试 TAP KL 算法的聚类性能,即芮氏指标(Rand Index, RI)^[21] 与归一化互信息(Normalized Mutual Information, NMI)^[19]。RI 和 NMI 的值均在[0,1]区间内,且其值 越接近 1, 算法所得的聚类性能越好。由于人脸数 据集呈现出明显的非线性流形结果,且同一个类别 的人脸图像,在不同的光照条件以及面部表情时, 能够拥有很高的相似性。因此,本文采用如表2所 示的 Extend Yale B, Yale 和 Olivetti 3 个人脸数据 集作为实验数据集^[22,23]。另一方面,为了从不同类 型数据集验证算法的有效性,本文还采用了 MNIST 手写体数据^[24]。对于 Extend Yale B 和 MNIST 数据 集,随机选取部分样本作为源域,剩余的少部分样 本构成目标域;对于样本量不充分的Yale与Olivetti 数据集,实验中通过图片顺时针旋转5°、逆时针旋 转5°、缩放0.9倍、缩放1.1倍,人为构造源域与目 标域数据集^[7,25]。为了更准确的分析比较各类聚类算 法的聚类性能,本节中的实验结果均是随机进行 30 次所得。

表 2 各数据集描述

数据集	样本 个数	类别数	样本 大小	源域样 本数	目标 域样 本数
Extend	2414	38	32×32	2204	210
Yale B					
MNIST	1000	10	28×28	800	200
Yale	165	15	70×70	165	30
Olivetti	400	40	$64{\times}64$	400	50

4.2 实验结果分析

需要说明的是,与文献[19]中的实验部分一致, 针对 Yale 和 Olivetti 数据集,首先对图像进行高斯 核滤波处理,高斯核参数为 0.5,其次对图像进行均 值为 0,方差为 0.1 的归一化处理。TAP_KL 算法 和 AP 算法中的偏向参数α的设置将会影响算法所 得的簇总数, α越大, 算法所得的簇总数越少, 反 之, 越小的α将导致越多的簇总数。综合数据集的 真实类标以及文献[13]中偏向参数α的设置办法,本 节中实验参数如表3所示。另外, TSC 算法要求预 先设定算法所得的簇总数,以及其他若干参数, 由 于篇幅原因,表3并没有标识出 TSC 算法的所有参 数,涉及的有关参数的设置均遵循文献[8]。

针对正则化参数 λ 对 TAP_KL 算法聚类性能 的影响,目标函数式(17)中第 1,第 2 项拥有同样的 量纲,因此 λ 的取值不必太大, λ 可以在{1,2,3,4,5, 6,7,8,9,10}范围内进行网格寻优。表 4 中列出了 λ ∈ {1,3,5,7,10}时,基于各数据集的 TAP_KL 算法的 聚类结果,并从平均值及标准差的角度进行说明。 分析表中数据可知, λ 从{1,2,3,4,5,6,7,8,9,10}范围 内进行网格寻优是可靠的。

为了公正、准确地分析比较各类聚类算法,实验中的各类算法都在数据集中运行多次,借助多个 聚类评价指标,并通过 t 检验(t-test)统计分析各类 算法在不同数据集上的实验结果。基于 4 个数据集 的各类算法的聚类结果具体如表 5-表 7 所示,以 TAP_KL 算法为基准,采用 t 检验统计分析各算法 的聚类性能,所得结果如表 8 所示,其中所有参数 设置均如表 3 所示。值得指出的是,当表 8 中的 *p* 值小于 0.05 时,统计学中认为成对比较的两类实验 结果具有显著性不同。由于 Olivetti 数据集的高维 度以及高类别数,基于 Olivetti 数据集的高维 运行时间过长,因此本文未将其与 TAP_KL 算法 比较。分析各算法的聚类性能,所得实验结论如下:

(1)分析表 5-表 7,考虑到 TAP_KL 算法在聚 类性能 RI 与 NMI 的表现,在源域样本数充分而目 标域样本数不足的情况下,TAP_KL 算法能够借助 源域的聚类结果完成目标域的聚类任务,并得到可 靠的聚类结果。

(2)由于目标域的数据集不充分,原始 AP 算法 所得的聚类性能低于迁移聚类算法的性能,尤其是 基于人脸数据集的实验中,当λ取值合理时,根据 表 8 中 TAP KL 算法与 AP 算法的 t 检验统计分析

表3 参数设置

数据集	Extend Yale B	Yale	Olivetti	MNIST	涉及算法
正则化参数 λ	$\lambda \in \{1,10\}$	$\lambda \in \{1,10\}$	$\lambda \in \{1,5\}$	$\lambda \in \{1,10\}$	TAP_KL
阻尼系数 l_d	$l_d = 0.5$	$l_d=0.5$	$l_{d} = 0.5$	$l_d=0.5$	TAP_KL, AP
偏向参数 α	欧氏距离的中值	$\alpha = 5$	$\alpha = 8$	欧氏距离的中值	TAP_KL, AP
预设簇总数	38	40	15	10	TSC^*

	表 4 TAP_KL 算法在不同 λ 时的聚类结果											
		数据集	Extend	Yale B	MN	IST		Yale	le Olivetti			
		聚类指标	均值	标准差	均值	标准差	均值	标准差	均值	标准差		
	1	RI	0.9529	0.0036	0.8660	0.0036	0.9244	0.0166	0.7349	0.0530		
	1	NMI	0.5756	0.0249	0.5534	0.0209	0.8037	0.0451	0.4934	0.0540		
正		RI	0.9545	0.0035	0.8662	0.0037	0.9271	0.0248	0.7786	0.0451		
	3	NMI	0.5844	0.0221	0.5485	0.0275	0.8111	0.0632	0.5396	0.0697		
化	5	RI	0.9533	0.0033	0.8677	0.0034	0.9428	0.0117	0.7839	0.0527		
参	0	NMI	0.5761	0.0226	0.5594	0.0290	0.8439	0.0316	0.5811	0.0615		
鉯	7	RI	0.9534	0.0023	0.8654	0.0030	0.9347	0.0155	0.8035	0.0481		
λ	1	NMI	0.5778	0.0176	0.5456	0.0200	0.8289	0.0328	0.5969	0.0769		
	10	RI	0.9550	0.0029	0.8676	0.0033	0.9306	0.0117	0.8492	0.0321		
_	10	NMI	0.5895	0.0227	0.5592	0.0193	0.8297	0.0248	0.6549	0.0537		

表 5 各算法基于数据集聚类性能比较

数据集			Extend	Yale B		MNIST			
聚类	性能	Ν	MI	F	RI		NMI RI		NI II
算法	参数	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差
TAP_KL -) 1	0.5271	0.6221	0.9466	0.9597	0.5112	0.5858	0.8602	0.8719
	$\lambda = 1$	0.5806	0.0225	0.9536	0.0035	0.5523	0.0208	0.8660	0.0029
	$\lambda = 10$	0.5644	0.6203	0.9511	0.9590	0.4966	0.6021	0.8603	0.8731
		0.5895	0.0227	0.9550	0.0029	0.5592	0.0193	0.8676	0.0039
TSC		0.3556	0.3760	0.9502	0.9531	0.5047	0.5601	0.8663	0.8839
		0.3671	0.0060	0.9519	0.0010	0.5347	0.0178	0.8736	0.0054
٨D		0.4817	0.5810	0.9339	0.9529	0.5108	0.6096	0.8587	0.8750
АГ		0.5226	0.0242	0.9432	0.0044	0.5543	0.0253	0.8660	0.0042

表 6 各算法基于 Yale 数据集聚类性能

		顺时针	旋转 5°	逆时针旋转 5°		缩放 0.9 倍		缩放 1.1 倍		
评价措	标				NI	II				
算法	参数	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	
TAP_KL		0.6873	0.9228	0.6907	0.9001	0.6599	0.8565	0.5931	0.8311	
	$\lambda = 1$	0.8066	0.0564	0.7950	0.0493	0.7792	0.0439	0.7223	0.0642	
		0.6838	0.9182	0.6810	0.8753	0.7012	0.8888	0.6431	0.8517	
	$\lambda = 5$	0.8386	0.0491	0.8029	0.0520	0.7975	0.0558	0.7443	0.0522	
TRO		0.5819	0.6502	0.6106	0.6645	0.5797	0.6633	0.5950	0.6666	
150		0.6240	0.0180	0.6350	0.0122	0.6202	0.0197	0.6291	0.0168	
A D		0.6806	0.9030	0.6966	0.8918	0.6599	0.8493	0.5931	0.8342	
AP		0.7918	0.0531	0.7975	0.0453	0.7721	0.0449	0.7192	0.0661	
评价措	标	RI								
算法	参数	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	
		0.8690	0.9724	0.8713	0.9609	0.8529	0.9471	0.7816	0.9333	
	$\lambda = 1$	0.9235	0.0253	0.9170	0.0248	0.9099	0.0251	0.8770	0.0374	
TAP_KL		0.8690	0.9724	0.8644	0.9586	0.8644	0.9632	0.8276	0.9448	
	$\lambda = 5$	0.9378	0.0207	0.9210	0.0248	0.9175	0.0264	0.8903	0.0298	
TSC		0.8980	0.9221	0.9039	0.9216	0.8934	0.9219	0.9035	0.9224	
150		0.9104	0.0054	0.9130	0.0045	0.9097	0.0057	0.9120	0.0048	
٨D		0.8713	0.9609	0.8690	0.9609	0.8437	0.9402	0.7839	0.9379	
Ar		0.9175	0.0221	0.9190	0.0227	0.9041	0.0278	0.8746	0.0393	

		顺时针旋转 5° 逆时针旋转 5°		缩放 0.9 倍		缩放	1.1 倍			
评价指标		NMI								
算法	参数	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	
		0.4233	0.7667	0.4081	0.6757	0.3489	0.6963	0.3489	0.6963	
TAD KI	$\lambda = 1$	0.5443	0.0832	0.5439	0.0695	0.5637	0.0816	0.5637	0.0816	
TAP_KL) 10	0.5734	0.7358	0.5594	0.7443	0.6002	0.8709	0.6002	0.8709	
	$\lambda = 10$	0.6591	0.0488	0.6624	0.0551	0.7420	0.0688	0.7420	0.0688	
٨D		0.4179	0.7450	0.4081	0.6757	0.3384	0.6925	0.3384	0.6925	
Ar		0.5309	0.0843	0.5170	0.0723	0.5119	0.0825	0.5119	0.0825	
评价打	指标	RI								
算法	参数	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	最小值 均值	最大值 标准差	
		0.6678	0.9151	0.6612	0.8784	0.5935	0.8767	0.5935	0.8767	
TAD KI	$\lambda = 1$	0.7772	0.0580	0.7828	0.0530	0.7853	0.0691	0.7853	0.0691	
IAF_KL) 10	0.7927	0.9020	0.7992	0.9118	0.8261	0.9567	0.8261	0.9567	
	$\lambda = 10$	0.8503	0.0317	0.8562	0.0331	0.8944	0.0369	0.8944	0.0369	
۸D		0.6678	0.9037	0.6612	0.8784	0.4294	0.8767	0.4294	0.8767	
Аг		0.7665	0.0607	0.7659	0.0540	0.7444	0.0869	0.7444	0.0869	

表7 各算法基于 Olivetti 数据集的聚类性能

表 8 各数据集 t 检验结果

参数	涉及算法	聚类性能	数据集						
			Extend	Yale B	M	NIST			
	TAD VI HAD	NMI	1.37H	E-17*	0.5253				
$\lambda = 1$	IAF_KL –JAF	RI	3.11H	E-16*	0.	8745			
$\lambda = 1$	TAD KI E TSC	NMI	7.48H	E-10*	0.1256				
	1AF_KL –J 150	RI	0.3	859	0.	0020*			
		NMI	3.17H	2-04*	0.	8053			
) 10	TAF_KL → AF	RI	4.76H	2-04*	0.	2463			
$\lambda = 10$	TAD KI STSC	NMI	1.18H	E-07*	0.	0403*			
		RI	0.02	242*	0.	0027*			
				YAL	Æ				
			顺时针旋转 5°	逆时针旋转 5°	缩放 0.9 倍	缩放 1.1 倍			
) - 1	TAP_KL 与 AP	NMI	0.0092*	0.4591	0.0580	0.2991			
		RI	0.0179^{*}	0.2573	0.0119^{*}	0.2044			
$\lambda = 1$	TAD KI E TSC	NMI	$3.32E{-}16*$	3.23E-16*	$4.25E-17^{*}$	7.38E-08*			
		RI	0.0074^{*}	0.4106	0.9621	$2.31\mathrm{E}\text{-}05^{*}$			
	TAP KI SAP	NMI	0.0012^{*}	0.7046	0.0707	0.1175			
) 10		RI	0.0008^{*}	0.7711	0.0838	0.0857			
$\lambda = 10$	TAD KI STSC	NMI	1.02E-21*	$9.73E{-}17*$	$3.14\mathrm{E}16^{*}$	$5.76E{-}14*$			
		RI	1.89E-08*	0.0975	0.1219	0.0003^{*}			
			Olivetti						
			顺时针旋转 5°	逆时针旋转 5°	缩放 0.9 倍	缩放 1.1 倍			
) 1		NMI	0.0092*	0.0016^{*}	$1.89E-05^{*}$	$1.89E{-}05^{*}$			
$\lambda = 1$	IAF_KL –JAF	RI	0.0308*	0.0027*	0.0004*	0.0004*			
) 10	ΤΔΡ ΚΙ Ξ ΔΡ	NMI	1.12E-06*	1.59E-08*	$3.89E-12^{*}$	$3.89E-12^{*}$			
$\lambda = 10$		RI	$3.47E-07^{*}$	3.36E-08*	$8.88E{-}10*$	$8.88E{-}10*$			

注:*标识出 p 值小于 0.05

结果,TAP_KL 算法的性能完全优于原始的 AP 算法;

(3)在与 TSC 算法的比较中,值得指出的是, TSC 算法需要预设数据集的簇总数。在这个前提下, 表 8 中 TAP_KL 算法与 TSC 算法的 t 检验统计分 析结果显示,在解决本节中涉及到的 4 个数据集的 迁移聚类问题时,TAP_KL 算法的性能优于 TSC 算法。

5 结束语

本文针对迁移聚类问题,提出了一种新的基于 Kullback-Leiber 距离的迁移仿射聚类算法,即 TAP KL 算法。相对于其他聚类算法,本文研究了 基础的 AP 算法,并进行改进以解决迁移聚类问题。 TAP KL 算法首先从概率角度重新解释 AP 算法的 目标函数。其次借助于信息论中的 KL 距离,测量 源域与目标域代表点集合的相似性。最后,通过详 细分析 TAP KL 算法与 AP 算法的目标函数,得出 一个重要结论,即可以将源域与目标域的相似性嵌 入到目标域数据集相似性矩阵的计算中,从而直接 利用 AP 算法的优化算法解决新的迁移聚类问题。 仿真实验分析进一步验证了 TAP KL 算法在解决 迁移聚类问题时的有效性。虽然本文所提算法在解 决迁移聚类问题时体现了较高的可靠性,算法仍然 存在一些需要解决的问题。例如,算法的聚类性能 并不十分稳定,在多次重复的随机实验中出现了一 定的差别及较高的标准差,如何提高算法的稳定性 是一个非常重要的工作,我们将在未来的工作中做 更深入的研究。另外,信息论中存在多种距离度量 方法,其他距离是否可以度量源域与目标域的相似 性,并进而发展出一种新的迁移聚类算法也是我们 在以后的工作中会关注的方向。

参考文献

- LONG M, WANG J, DING G, et al. Adaptation regularization: A general framework for transfer learning[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(5): 1076–1089. doi: 10.1109/TKDE.2013.111.
- [2] 毕安琪,王士同.基于 SVC 和 SVR 约束组合的迁移学习分类
 算法[J].控制与决策,2014,29(6):1021-1026.doi: 10.13195
 /j.kzyjc.2013.0520.

BI Anqi and WANG Shitong. Transfer classification learning based on combination of both SVC and SVR's constraints[J]. *Control and Decision*, 2014, 29(6): 1021–1026. doi: 10.13195/j. kzyjc.2013.0520.

[3] PATRICIA N and CAPUTO B. Learning to learn, from transfer learning to domain adaptation: A unifying perspective[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014: 1442–1449. doi: 10.1109/CVPR.2014.187.

- [4] XU Z J and SUN S L. Multi-source transfer learning with multi-view adaboost[J]. Neural Information Processing, 2012, 7665: 332–339. doi: 10.1007/978-3-642-34487-9_41.
- [5] PAN S J L, KWOK J T, and YANG Q. Transfer learning via dimensionality reduction[C]. Proceedings of the 23rd International Conference on Artificial Intelligence, CA, USA: 2008: 677–682.
- [6] PAN S J L, NI X C, SUN J T, et al. Cross domain sentiment classification via spectral feature alignment[C]. Proceedings of the 19th International Conference on World Wide Web (WWW-10). New York, USA, 2010, 751–760. doi: 10. 1145/1772690.1772767.
- [7] 蒋亦樟,邓赵红,王士同. ML 型迁移学习模糊系统[J]. 自动 化学报, 2012, 38(9): 1393-1409. doi: 10.3724/SP.J.1004.2012. 01393.

JIANG Yizhang, DENG Zhaohong, and WANG Shitong. Mamdani-Larsen type transfer learning fuzzy system[J]. Acta Automatica Sinica, 2012, 38(9): 1393–1409. doi: 10. 3724/SP.J.1004.2012.01393.

- [8] JIANG W H and CHUNG F L. Transfer spectral clustering [C]. Proceedings of the 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Berlin, Heidelberg, 2012: 789–803. doi: 10.1007/978-3-642-33486-3_50.
- [9] LI M J, NG M K, CHEUNG Y M, et al. Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2008, 20(11): 1519–1534. doi: 10.1109/TKDE. 2008.88.
- [10] KRISHMA K and MURTY M N. Genetic K-means algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 1999, 29(3): 433–439. doi: 10.1109/3477. 764879.
- [11] ERSAHIN K, CUMMING I G, and WARD R K. Segmentation and classification of Polari metric SAR data using spectral graph partitioning[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2010, 48(1): 164–167. doi: 10.1109/TGRS.2009.2024303.
- [12] LAUER F and SCHNORR C. Spectral clustering of linear subspaces for motion segmentation[C]. Proceedings of the 12th IEEE International Conference of Computer Vision, Kyoto, Japan, 2009: 678–685. doi: 10.1109/ICCV. 2009. 5459173.
- [13] FREY B J and DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972–976.
- [14] 肖宇, 于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报, 2008, 19(11): 2803-2813. doi: 10.3724/SP.J.1001.2008.02803.

- [15] 储岳中, 徐波, 高有涛. 基于近邻传播聚类与核匹配追踪的遥 感图像目标识别方法[J]. 电子与信息学报, 2014, 36(12):
 2923-2928. doi: 10.3724/SP.J.1146.2014.00422.
 CHU Yuezhong, XU Bo, and GAO Youtao. Technique of remote sensing image target recognition based on affinity propagation and kernel matching pursuit[J]. Journal of Electronics & Information Technology, 2014, 36(12):
 2923-2928. doi: 10.3724/SP.J.1146.2014.00422.
- [16] SUN L and GUO C H. Incremental affinity propagation clustering based on message passing[J]. *IEEE Transactions* on Knowledge and Data Engineering, 2014, 26(11): 2731–2744. doi: 10.1109/TKDE.2014.2310215.
- [17] SHI X H, GUAN R C, WANG L P, et al. An incremental affinity propagation algorithm and its applications for text clustering[C]. Proceedings International Joint Conference on Neural Networks, Atlanta, GA, USA, 2009: 2914–2919. doi: 10.1109/IJCNN.2009.5178973.
- [18] ZHENG Yun and CHEN Pei. Clustering based on enhanced α-expansion move[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 25(10): 2206–2216. doi: 10.1109/ TKDE.2012.202.
- [19] 毕安琪,董爱美,王士同.基于概率和代表点的数据流动态聚 类算法[J]. 计算机研究与发展, 2016, 53(5): 1029-1042.
 BI Anqi, DONG Aimei, and WANG Shitong. A dynamic data stream clustering algorithm based on probability and exemplar[J]. Journal of Computer Research and Development, 2016, 53(5): 1029-1042.
- [20] 孙力娟,陈小东,韩崇,等. 一种新的数据流模糊聚类方法[J].
 电子与信息学报,2015,37(7):1620-1625. doi: 10.11999/ JEIT141415.

SUN Lijuan, CHEN Xiaodong, HAN Chong, et al. New fuzzyclustering algorithm for data stream[J]. Journal of Electronics & Information Technology, 2015, 37(7): 1620–1625. doi: 10.11999/JEIT141415.

- [21] JIANG Y Z, CHUNG F L, WANG S T, et al. Collaborative fuzzy clustering from multiple weighted views[J]. *IEEE Transactions on Cybernetics*, 2015, 45(4): 688–701. doi: 10. 1109/TCYB.2014.2334595.
- [22] CAI D, HE X F, HAN J W, et al. Orthogonal Laplacian faces for face recognition[J]. *IEEE Transactions on Image Processing*, 2006, 15(11): 3608–3614. doi: 10.1109/TIP.2006. 881945.
- [23] 张景祥,王士同,邓赵红,等.融合异构特征的子空间迁移学 习算法[J]. 自动化学报,2014,40(2):236-246.doi: 10.3724/SP.J. 1004.2014.00236.

ZHANG Jingxiang, WANG Shitong, DENG Zhaohong, et al. A subspace transfer learning algorithm integrating heterogeneous features[J]. Acta Automatica Sinica, 2014, 40(2): 236–246. doi: 10.3724/SP.J.1004.2014.00236.

- [24] LE C Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324. doi: 10.1109/5.726791.
- [25] REN J, SHI X, FAN W, et al. Type independent correction of sample selection bias via structural discovery and rebalancing[C]. Proceedings of the 8th SIAM International Conference on Data Mining, Atlanta, GA, USA, 2008: 565–576. doi: 10.1137/1.9781611972788.52.
- 毕安琪: 女,1989年生,博士生,研究方向为模式识别、人工智能、迁移学习等.
- 王士同: 男,1964年生,教授,研究方向为模式识别、人工智能、 模糊系统等.