文章编号: 1001-0920 (2015) 06-1000-07 DOI: 10.13195/j.kzyjc.2014.0770

知识迁移极大熵聚类算法

钱鹏江, 孙寿伟, 蒋亦樟, 王士同, 邓赵红

(江南大学 数字媒体学院，江苏 无锡 214122)

摘 要: 为解决数据不足或失真等环境下传统聚类技术效果不佳的问题, 基于历史类中心和历史隶属度提出两种知识迁移机制, 并与极大熵聚类方法融合提出知识迁移极大熵聚类算法 KT-MEC. KT-MEC的优点是: 利用历史知识, KT-MEC 聚类有效性和实用性明显增强; 内嵌迁移机制均不暴露源域数据, 从而拥有源域隐私保护能力; KT-MEC 基于的“参数寻优+聚类有效性度量”机制理论上保证其性能不差于经典极大熵算法, 避免了负迁移问题.

关键词: 知识迁移；极大熵聚类；隐私保护；负迁移

中图分类号: TP391.4 **文献标志码:** A

Knowledge transfer based maximum entropy clustering

QIAN Peng-jiang, *SUN Shou-wei*, *JIANG Yi-zhang*, *WANG Shi-tong*, *DENG Zhao-hong*

(School of Digital Media，Jiangnan University，Wuxi 214122，China．Correspondent：QIAN Peng-jiang，E-mail: qpengjiang@gmail.com)

Abstract: Classical clustering methods tend to be less effective in such situation where the data are insufficient or impure. Therefore, two knowledge transfer mechanisms for fuzzy partition clustering are devised in terms of historical cluster centers and fuzzy memberships regarding historical class centers respectively. And combining these two transfer mechanisms with the classical maximum entropy clustering(MEC) approach, the particular knowledge transfer based maximum entropy clustering(KT-MEC) algorithm is proposed. The major merits of KT-MEC lie in following three aspects. Benefiting from the auxiliary guidance of historical knowledge, the clustering effectiveness and practicability of KT-MEC are enhanced distinctly. As the couple of built-in transfer mechanisms both don't expose the raw data in the source domain, KT-MEC is of good capability of privacy protection for the source domain. Owing to the "searching for best parameters + validity indices" mechanism, the clustering effectiveness of KT-MEC is not worse than that of MEC in theory, which avoids reliably the negative transfer risk.

Keywords: knowledge transfer；maximum entropy clustering；privacy protection；negative transfer learning

0 引 言

传统聚类方法一般以大量的可用数据为基础进行信息挖掘、模型学习和实践验证. 在数据积累初期, 信息往往有限甚至匮乏, 且数据受污染的情况普遍存在, 这便给传统机器学习带来了新挑战.

研究人员发现, 迁移学习能较好地解决如数据量少、信息缺失和数据失真等现象[1-8]. 关于迁移学习, 目前已有的研究较多集中在分类学习[1-2]领域, 在模式识别其他领域, 基于迁移学习的理论或方法相对较少. 文献 [3-4] 进行了迁移回归模型的研究; 文献 [7-8] 进行了迁移聚类方法的探索, 本文着眼点正是基于迁移学习的新型聚类方法研究.

划分聚类是最常见的聚类方法之一. 模糊 C 均值 (FCM) 聚类算法[9-11]、极大熵聚类 (MEC) 算法[12-13]等是其中的典型代表. MEC 算法以简洁的数学表达和明确的物理含义引起许多研究人员的兴趣, 如文献 [14] 尝试提高 MEC 对异常点的识别能力, 文献 [15] 提出的基于模糊线性判别分析的极大熵模糊聚类算法等.

本文进行了基于划分聚类的迁移学习问题研究, 首先提出两种适用的学习机制: 1) 基于历史类中心的知识迁移机制, 该迁移机制通过调控源域和目标域类

收稿日期: 2014-05-17；**修回日期:** 2014-08-06.

基金项目: 国家自然科学基金项目(61202311)；江苏省自然科学基金项目(BK201221834)；江苏省产学研前瞻性研究项目(BY2013015-02).

作者简介: 钱鹏江(1979−), 男, 副教授, 博士, 从事模式识别、图像处理等研究；孙寿伟(1989−), 男, 硕士生, 从事智能算法及应用的研究.

中心的一致性程度达到迁移学习的效果; 2) 目标域数据相对于历史类中心的隶属度的迁移机制, 该机制分别计算目标域数据点相对于源域历史类中心和目标域当前估计类中心的隶属度, 使这两种隶属度通过平衡因子相组合达到迁移学习的目的. 将上述所提出的双重迁移机制融入经典的MEC算法, 提出了知识迁移极大熵聚类算法(KT-MEC), 该算法具备以下特色:

1) 由于迁移知识同时基于历史类中心和目标域数据相对于历史类中心的隶属度, 迁移学习的能力得到保证, 有效增强了KT-MEC的聚类性能和实用性.

2) KT-MEC的参数寻优机制结合有效的聚类有效性度量理论上覆盖了经典MEC的情况, 即后者是前者的一种特例, 这保证KTMEC最终性能较原MEC算法只会提升不会减弱, 避免了所谓的负迁移问题.

3) 本文提出的知识迁移机制, 无论基于历史类中心还是基于关于历史类中心的隶属度, 均不暴露源域原始数据, 这样KT-MEC算法具备较好的源域隐私保护能力.

1　极大熵聚类算法

MEC算法以独特的熵概念重构C均值算法的目标函数, 得到了最大熵意义下的模糊聚类算法. 经典MEC算法具体表述[11]如下: 给定样本空间 $\boldsymbol{X}=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in R^d,\ i=1,2,\cdots,N\}$, 内含 $C\,(2\leqslant C<N)$ 个不同类, 则经典MEC算法目标函数式为

$$\begin{aligned}&J_{\mathrm{MEC}}(\boldsymbol{U},\boldsymbol{V})=\\&\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\gamma\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\ln\mu_{ij};\\&\text{s.t. }\mu_{ij}\in[0,1],\ 1\leqslant i\leqslant C,\ 1\leqslant j\leqslant N,\ \sum_{i=1}^{C}\mu_{ij}=1.\end{aligned}\tag{1}$$

其中: C 为类别总数, N 为样本总数, γ 通常称为正则化参数, $\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2$ 为第 j 个样本与第 i 个类中心之间的距离, μ_{ij} 为第 j 个样本相对于第 i 个类中心的隶属度, $\boldsymbol{U}$ 为由 μ_{ij} 构成的隶属度矩阵 $\boldsymbol{U}\in R^{N\times C}$, $\boldsymbol{v}_i$ 为第 i 类的类中心, $\boldsymbol{V}$ 为由 $\boldsymbol{v}_i$ 组成的类中心矩阵.

根据拉格朗日条件极值最优化方法, 求解式(1)得到最优解时中心点 $\boldsymbol{V}$ 和隶属度 $\boldsymbol{U}$ 的迭代公式为

$$\boldsymbol{v}_i=\sum_{j=1}^{N}\mu_{ij}\boldsymbol{x}_j\Big/\sum_{j=1}^{N}\mu_{ij},\ i=1,2,\cdots,C;\tag{2}$$

$$\begin{aligned}&\mu_{ij}=\frac{\exp\Big(-\dfrac{\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2}{\gamma}\Big)}{\displaystyle\sum_{k=1}^{C}\exp\Big(-\dfrac{\|\boldsymbol{x}_j-\boldsymbol{v}_k\|^2}{\gamma}\Big)},\\&i=1,2,\cdots,C,\ j=1,2,\cdots,N.\end{aligned}\tag{3}$$

2　模糊划分聚类的两种知识迁移学习机制

迁移学习(TL)是从现有场景(又称为源域)获取有用信息, 用于指导与现有场景具有一定相关性但又存在明显差异的别的场景(目标域)中的学习过程的机器学习模式, 其中从源域获取的有用信息存在源数据和知识两种形式, 前者是迁移学习的最基本形式, 如带类标的部分代表点; 后者是迁移学习的高级形式, 可有效避免负迁移问题.

在基于划分的模糊聚类算法中, 类中心 $\boldsymbol{V}$ 和隶属度 $\boldsymbol{U}$ 作为决定聚类最终结果的两个重要因素, 拥有丰富信息, 是一种潜在的可用知识. 以类中心 $\boldsymbol{V}$ 和隶属度 $\boldsymbol{U}$ 作为知识用于进行跨领域模糊划分聚类问题研究, 提出两种适于模糊划分聚类的知识迁移机制.

1) 基于历史类中心的知识迁移为

$$\Phi(\boldsymbol{V},\tilde{\boldsymbol{V}})=\lambda\sum_{i=1}^{C}\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2.\tag{4}$$

其中: $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_C]$ 为目标域估计的类中心矩阵; $\tilde{V}=[\tilde{v}_1,\tilde{v}_2,\cdots,\tilde{v}_C]$ 为源域历史类中心矩阵, 其作为类中心知识用于迁移学习, 可通过经典MEC算法对源域数据集进行聚类分析获得.

式(4)将源域类中心 $\tilde{v}_i\,(i=1,2,\cdots,C)$ 作为参考进行迁移学习, 确保在目标函数取最优时, 目标域的类中心在一定程度上与源域历史类中心一致. 当 $\lambda\geqslant 0$ 时, 为正则化参数; 当 $\lambda\to 0$ 时, 表示历史类中心知识 $\tilde{V}$ 不可靠、可借鉴度低; 当 $\lambda\to+\infty$ 时, 表示已知的历史类中心知识 $\tilde{V}$ 有效, 可以极好地指导目标域的聚类工作.

2) 基于目标域数据相对于历史类中心的隶属度的知识迁移为

$$\Theta(\boldsymbol{U},\tilde{\boldsymbol{U}})=\sum_{i=1}^{C}\sum_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})d_i.\tag{5}$$

其中: d_i 为目标域中与类中心 $\boldsymbol{v}_i$ 相关的某距离度量(如某数据点 $\boldsymbol{x}_j$ 相对于 $\boldsymbol{v}_i$ 的距离); μ_{ij} 为目标域内数据点 $\boldsymbol{x}_j$ 相对于目标域当前估计类中心 $\boldsymbol{v}_i$ 的隶属度; $\tilde{\mu}_{ij}$ 为目标域内数据点 $\boldsymbol{x}_j$ 相对于源域历史类中心 $\tilde{v}_i$ 的隶属度, 可由式(3)获得. 式(5)将目标域数据相对于源域类中心的隶属度和相对于目标域估计类中心的隶属度相融合, 以进行跨域知识学习, 其基本思想是目标域估计的类中心应该与源域已知的历史类中心存在一定程度的相似性, 这正是跨域迁移学习的基本前提. 平衡因子 $0\leqslant\eta\leqslant 1$ 决定两者的重要程度, $\eta\to 1$ 表示强调相对于目标域当前类中心的隶属度的作用, 反之表示强调相对于源域历史类中心的隶属度信息.

式(5)表达的是一种加权求和的内涵, 每个类

中心的权重由$\sum\limits_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})$决定. 属于某类的数据点越多, $\sum\limits_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})$便越大, 从而$\sum\limits_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})d_i$在整个式(5)中占的比重就越大, 对整个表达式的影响程度也越大.

上述两种迁移机制均是高级知识的形式, 具有高度的数据抽象特征, 因此使用它们作为迁移知识不会暴露源域的原始数据内容, 对源域数据起到了较好的隐私保护作用.

3 知识迁移极大熵聚类算法

3.1 历史类中心和基于历史类中心的隶属度的同步迁移MEC目标函数

为KT-MEC构造如下融入双重知识迁移学习机制的特定极大熵模糊划分聚类架构, 即历史类中心与基于历史类中心的隶属度同步迁移模式, 其具体目标函数为

$$\begin{aligned}&J_{\mathrm{KT-MEC}}(\boldsymbol{U},\boldsymbol{V})=\\&\sum_{i=1}^{C}\sum_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\\&\gamma\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\ln\mu_{ij}+\\&\lambda\sum_{i=1}^{C}\sum_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2;\\&\text{s.t. }\mu_{ij}\in[0,1],\ 1\leqslant i\leqslant C,\ 1\leqslant j\leqslant N,\ \sum_{i=1}^{C}\mu_{ij}=1.\end{aligned}\tag{6}$$

其中: γ和λ分别为熵和知识迁移正则化系数, η为迁移平衡因子, $\gamma>0$, $\lambda\geqslant 0$, $\lambda\neq 1$, $0\leqslant\eta\leqslant 1$. 式(6)其余各成员含义同第1节和第2节所述. 注意到, 当参数$\lambda=0$且$\eta=1$时, KT-MEC算法实际上退化为经典MEC算法, 这为本文有效避免负迁移提供了保障.

该架构的含义是借鉴历史类中心和基于历史类中心的隶属度的双重迁移知识, 寻求式(6)的最优解. 式(6)目标函数第1项是式(5)的具体形式, 即借鉴部分相对于历史类中心的隶属度信息计算目标域各数据点相对于当前类中心的加权距离总和; 第3项是式(4)和(5)的融合形式, 即同步借鉴历史类中心和基于历史类中心的隶属度信息约束目标域当前估计类中心较之历史类中心的偏离程度, 它也是一种基于加权和的度量, 即某类包含的数据点越多, 对整体目标函数的影响成分便越大.

3.2 KT-MEC算法的迭代公式

基于拉格朗日条件极值最优化方法可以获得KT-MEC算法的关于目标域类中心和隶属度的迭代公式, 具体推导过程如下: 首先构建Lagrange表达式

$$\begin{aligned}L=&\sum_{i=1}^{C}\sum_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\\&\gamma\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\ln\mu_{ij}+\\&\lambda\sum_{i=1}^{C}\sum_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2+\\&\sum_{j=1}^{N}\alpha_j\Big(1-\sum_{i=1}^{C}\mu_{ij}\Big),\end{aligned}\tag{7}$$

其中α_j为Lagrange乘子.

3.2.1 中心点$\boldsymbol{v}_i$的迭代公式

令$\partial L/\partial\boldsymbol{v}_i=0$, 解得

$$\boldsymbol{v}_i=\frac{\sum\limits_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})\boldsymbol{x}_j-\lambda\sum\limits_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})\tilde{\boldsymbol{v}}_i}{\sum\limits_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})-\lambda\sum\limits_{j=1}^{N}(\eta\mu_{ij}+(1-\eta)\tilde{\mu}_{ij})}.\tag{8}$$

其中: $\gamma>0$, $\lambda\geqslant 0$, $\lambda\neq 1$, $0\leqslant\eta\leqslant 1$.

3.2.2 隶属度μ_{ij}的迭代公式

令$\partial L/\partial\mu_{ij}=0$, 有

$$\begin{aligned}&\mu_{ij}=\\&\exp\Big(\frac{\alpha_j-\gamma-\eta\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2-\lambda\eta\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2}{\gamma}\Big)=\\&\exp\Big(\frac{\alpha_j-\gamma}{\gamma}\Big)\exp\Big(-\frac{\eta(\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\lambda\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2)}{\gamma}\Big).\end{aligned}\tag{9}$$

又因为

$$\begin{aligned}&\sum_{i=1}^{C}\mu_{ij}=\\&\sum_{i=1}^{C}\exp\Big(\frac{\alpha_j-\gamma-\eta(\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\lambda\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2)}{\gamma}\Big)=1,\end{aligned}\tag{10}$$

即

$$\exp((\alpha_j-\gamma)/\gamma)=\frac{1}{\sum\limits_{k=1}^{C}\exp\Big(-\frac{\eta(\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\lambda\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2)}{\gamma}\Big)},\tag{11}$$

将式(11)代入(9)可得

$$\mu_{ij}=\frac{\exp\left(-\dfrac{\eta(\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2+\lambda\|\boldsymbol{v}_i-\tilde{\boldsymbol{v}}_i\|^2)}{\gamma}\right)}{\sum\limits_{k=1}^{C}\exp\left(-\dfrac{\eta(\|\boldsymbol{x}_j-\boldsymbol{v}_k\|^2+\lambda\|\boldsymbol{v}_k-\tilde{\boldsymbol{v}}_k\|^2)}{\gamma}\right)}. \tag{12}$$

其中: $\gamma>0, \lambda\geqslant 0, \lambda\neq 1, 0\leqslant\eta\leqslant 1$.

根据迭代式 (8) 和 (12), 最终可求得目标域的最优类中心 $\boldsymbol{V}$ 和隶属度 $\boldsymbol{U}$.

3.3 KT-MEC 算法步骤

根据上述推导得到的迭代公式, 对 KT-MEC 算法进行具体描述如下.

输入: 类别数 C, 熵正则化参数 γ, 源域数据集 X_S, 目标域数据集 X_T, 迁移正则化参数 λ 和平衡因子 η, 迭代终止条件 ε, 最大迭代次数 f;

输出: 目标域类中心 $\boldsymbol{V}$ 和隶属度 $\boldsymbol{U}$.

源域知识总结阶段.

Step 1: 通过经典 MEC 算法对源域 X_S 进行分析, 得到历史类中心知识 $\tilde{V}$.

Step 2: 通过式 (3) 计算目标域 X_T 中数据对于源域历史类中心 $\tilde{V}$ 的隶属度知识 $\tilde{U}$.

目标域迁移学习阶段.

Step 3: 初始化迭代计数器 $t=0$, 随机初始化隶属度矩阵 $\boldsymbol{U}(t)$.

Step 4: 通过当前隶属度矩阵 $\boldsymbol{U}(t)$ 和历史隶属度知识 $\tilde{U}$, 由式 (8) 计算类中心矩阵 $\boldsymbol{V}(t+1)$.

Step 5: 通过 $\boldsymbol{V}(t+1)$ 和历史类中心 $\tilde{V}$, 由式 (12) 求隶属度矩阵 $\boldsymbol{U}(t+1)$.

Step 6: 当 $||\boldsymbol{U}(t+1)-\boldsymbol{U}(t)||_F<\varepsilon$ 或迭代次数 t 达到 f 时算法终止, 否则置 $t=t+1$, 返回 Step 4.

Step 7: 算法收敛后, 输出目标域类中心 $\boldsymbol{V}$ 和隶属度 $\boldsymbol{U}$.

KT-MEC 算法三核心参数 η, γ 和 λ 的自适应设置问题仍是尚待深入研究的课题. 本文策略是: 对于实验研究, 在目标域数据集类标已知的情况下, 借助 NMI[16,20]、RI[20]等权威外部有效性度量指标, 采用网格搜索策略获得算法的最佳参数设置. 对于无类标信息的真实数据集和实际应用场合, 如 DBI[17]、DI[17]等内部有效性度量指标也可以在一定程度上协助确定各参数的较好范围. 另外, 若用户提供目标域部分数据的分类情况供参考, 则常用的交叉验证方法也可用于本文参数值的设置.

4 实验分析

4.1 实验设置

为了验证 KT-MEC 算法的有效性, 分别基于一些人造和真实迁移场景数据集进行实验比较研究. 另外选择基于多任务学习机制的 LSSMTC 算法[18]、combKM 算法[18]、STC 算法[7]和 TSC 算法[8]参与实验比较. 实验采用的硬件配置为 Windows 7 64 Bit 和 8 GB 内存, 编程环境为 Matlab 2010a.

在与相关算法进行性能比较时, 采用两种评价指标进行分析讨论: 归一化互信息 (NMI)[16,20]和芮氏指标 (RI)[20]. 两种指标取值范围均为 [0,1], 数值越高表示聚类性能越好. 实验分别利用模拟数据集、文本数据集、人脸图像数据集对 KT-MEC 算法进行验证与评估.

在实验中, 有关参数设置采用网格搜索法进行遍历寻优. 4 个对比算法仍采用其原文献推荐的参数区间进行设置; MEC 算法参数 γ 的寻优区间在 $\{0.1:0.1:1\}\bigcup\{2:1:10\}\bigcup\{20:10:100\}$ 之间; KT-MEC 算法参数 λ、η 和 γ 的寻优区间分别为 $\{0,0.5\}\bigcup\{2:1:10\}\bigcup\{20:10:100\}$, $\{0:0.1:1\}$ 和 $\{0.1:0.2:1\}\bigcup\{2:1:10\}\bigcup\{20:10:150\}$. 实验结果中的数据均是运行 10 次后取得均值和方差的结果.

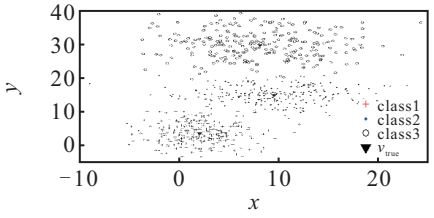
需要说明的是, TSC 算法要求数据维度要大于聚类的类目数, 因此对于无法满足该要求的数据集, 以“–”表示其无法正常执行.

4.2 结果分析

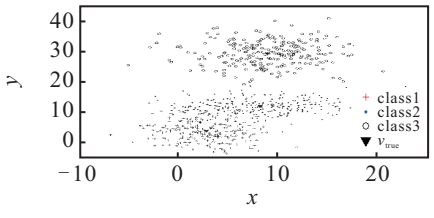
4.2.1 模拟数据实验和结果分析

采用高斯随机函数生成相关的模拟数据集, 类别数为 3, 维数为 2. 源域数据集和目标域数据集的生成过程如下:

1) 源域数据集 $\tilde{X}$ 和潜在理想目标域. 首先生成源域数据集, 如图 1(a) 所示, 其数据量相对充足, 包含 750 个样本点, 3 类, 每类包含 250 个样本点. 为了检验算法的性能, 给出潜在理想的目标域数据集, 如图 1(b) 所示. 之所以称为潜在, 是因为当前并不存在, 是目标域数据最终的分布趋势, 理想是假定其干净、未



(a) 源域数据集

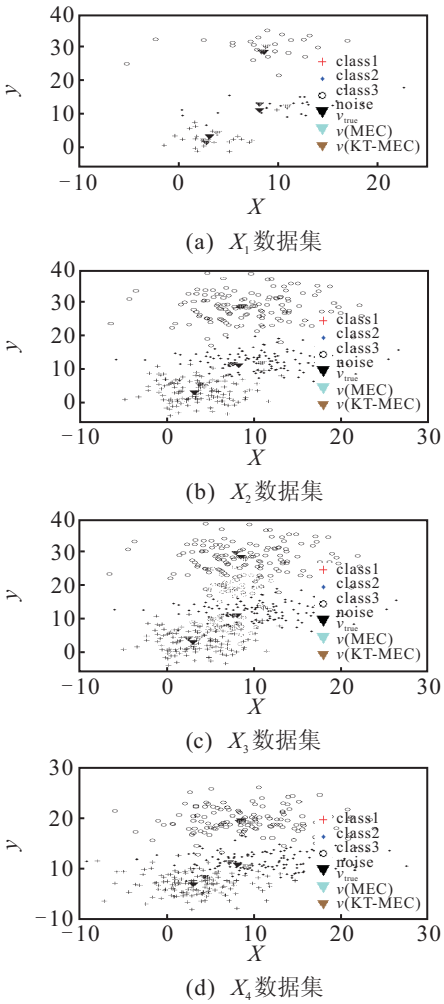


(b) 潜在理想目标域数据集

图 1　源域数据集 $\tilde{X}$ 和潜在理想目标域数据集

受污染. 如图1所示, 源域和潜在目标域数据分布类似但又有明显区别, 图1中, class$i$ ($i=1,2,3$) 代表第 i 类数据, V_{true} 代表各类的类中心.

2) 目标域数据集. 目标域数据由潜在理想目标域而派生, 它模拟了目标域此刻已有数据的分布情况, 是目标域不同时刻不同情形的缩影. 这里设计4种情况: ① 从潜在目标域中按类别等量抽取少量数据生成 X_1 数据集, 效果如图2(a)所示; ② 从潜在目标域中按类别等量抽取较大量数据生成 X_2 数据集, 效果如图2(b)所示; ③ 在 X_2 数据集中加两类干扰数据点构成数据集 X_3, 每类为35个干扰点, 如图2(c)所示; ④ 在 X_2 数据样本上加入均值为0、方差为2的高斯噪音构成数据集 X_4, 效果如图2(d)所示. class$i$ ($i=1,2,3$) 代表第 i 类数据, V_{true} 代表潜在的真实类中心.



(a) X_1数据集

(b) X_2数据集

(c) X_3数据集

(d) X_4数据集

图2 目标域数据集及各类中心分布情况

基于上述模拟迁移场景, 分别执行MEC、KT-MEC和另外4个对比算法, 并按 $\sum_{i=1}^{C}\|\boldsymbol{v}_i-\boldsymbol{v}_{\text{true}_i}\|^2$ 计算MEC和KT-MEC在各数据集上所得类中心与潜在真实类中心的距离, 得到如表1、表2和图2所示的实验结果. 图2中: v(MEC) 表示MEC算法在相应目标域数据集上聚得的类中心, v(KT-MEC) 表示KT-MEC算法聚得的类中心.

表1 6种算法在模拟数据集上的性能对比

数据集	评价指表	算法					
		LSSMTC	combKM	MEC	STC	TSC	KT-MEC
X_1	NMI-mean	0.762 6	0.785 9	0.813 0	0.855 2	–	**0.914 4**
	NMI-std	1.17e-16	9.79e-17	1.17e-16	0	–	0
	RI-mean	0.902 3	0.916 8	0.918 9	0.944 3	–	**0.965 4**
	RI-std	1.17e-16	2.34e-16	1.17e-16	0	–	1.17e-16
X_2	NMI-mean	0.820 5	0.824 3	0.826 0	0.855 2	–	**0.864 2**
	NMI-std	0	8.28e-17	1.74e-16	0	–	1.17e-16
	RI-mean	0.933 1	0.938 4	0.936 2	0.941 8	–	**0.949 2**
	RI-std	1.17e-16	1.17e-16	0	0	–	1.17e-16
X_3	NMI-mean	0.644 3	0.696 0	0.649 8	**0.733 5**	–	0.713 8
	NMI-std	1.17e-16	0.003 6	1.17e-16	0	–	1.17e-16
	RI-mean	0.838 2	0.873 8	0.840 5	0.882 2	–	**0.889 1**
	RI-std	1.17e-16	0.001 4	1.17e-16	0	–	1.17e-16
X_4	NMI-mean	0.726 3	0.716 3	0.752 3	0.778 2	–	**0.795 4**
	NMI-std	1.17e-16	1.17e-16	0	0	–	0
	RI-mean	0.898 4	0.891 7	0.911 1	0.916 8	–	**0.929 6**
	RI-std	0	2.34e-16	1.17e-16	0	–	0

表2 数据集上算法聚类中心与真实类中心的距离

算法	X_1	X_2	X_3	X_4
MEC	5.750 6	5.618 4	5.965 2	5.760 8
KT-MEC	3.794 9	3.342 8	2.875 4	4.844 8

观察表1和表2数据可得到如下结论:

1) 在数据量严重不足(即 X_1 数据集)时, 若继续使用经典的MEC算法, 则聚类效果明显不如KT-MEC, 因为数据量极其缺少时(如图2(a)所示), MEC算法根据仅有的数据所获取的类中心 v(MEC) 偏离了实际类中心 V_{true}. KT-MEC算法有效地参照了源域类中心和组合隶属度划分, 提高了最终的聚类效果.

2) 在目标域数据量较充足且未失真(即 X_2 数据集)时, 5个算法聚类效果均较为理想(人造数据场景不满足TSC执行的前提). 在 X_3 和 X_4 数据集, KT-MEC算法的优势较明显, 这是因为 X_3 和 X_4 数据存在一定失真, 普通算法难以有效地获得准确的类中心, 从而导致聚类效果较差. KT-MEC算法基于源域迁移知识, 获得了较好的聚类效果, 这表明了其具有较好的抗噪和抗干扰性能.

3) 在 X_3 上, STC聚类效果稍好于本文KT-MEC, 但STC工作时使用了完整的源域数据集, KT-MEC仅涉及源域类中心知识, 这意味着KT-MEC具有源域数据隐私保护能力, 而STC无此能力.

4) LSSMTC和CombKM是多任务聚类算法, 在工作时基于完整的源域数据集, 一方面它们缺乏源域数据隐私保护能力, 另一方面实验结果表明, 在此人造数据场景中, 这两个算法并不适合含有噪声或干扰数据的聚类任务.

5) 实验表明, KT-MEC较MEC所得的实际类中心更趋近潜在真实类中心.

4.2.2　文本数据集实验和结果分析

从 20Newsgroups(20NG)[20]数据库的 4 个大类中各选两个大类构成两个迁移场景“rec VS talk”和“comp VS sci”, 每个场景的每个大类又分别选择两个子类, 一个用于构成源域数据集, 另一个构成目标域数据集, 因此, 每个迁移场景的源域和目标域数据集也均包含两个子类, 如表 3 所示. 原始 20 NG 数据维数较高, 利用 BOW 工具箱[21]对其进行必要的降维处理, 两个迁移场景的数据具体构成如表 4 所示.

表 3　20 NG 中选择的数据集

数据类别	源域数据集	目标域数据集
Rec VS Talk	rec.autos talk.politics.guns	rec.sport.baseball talk.politics.mideast
comp VS sci	comp.sys.mac sci.med	comp.sys.ibm.pc sci.electronics

表 4　选用的 NG20 数据集说明

数据类别	数据集	样本个数	维数	类别
Rec VS Talk	源域数据	1500	350	2
	目标域数据	500	350	
comp VS sci	源域数据	1500	350	2
	目标域数据	500	350	

6 个对比算法在此两个文本迁移场景数据集上的执行结果如表 5 所示, 由表 5 可得到以下结论:

1) “rec VS talk”迁移场景数据集上, KT-MEC 算法优势明显, 这受益于历史类中心和基于历史类中心的隶属度知识的辅助学习过程. STC 算法和 TSC 算法作为两种另外的迁移聚类算法受限于所采用的迁移学习策略, 在此场景, 其迁移质量并不高, 聚类性能并不突出.

2) “comp VS sci”迁移场景数据集上, TSC 算法聚类效果最好, KT-MEC 算法次之. 究其原因, “comp VS sci”场景源域与目标域的相关性相对较弱, 即源域历史类中心及其相关的历史隶属度对目标域的指导性较差, 这点由 MEC 和 KT-MEC 的性能指标相近可得到证实. TSC 能够获得最好的聚类效果, 原因在于其使用的是完整的源域数据而非类中心知识, 因此受影响程度小于 KT-MEC.

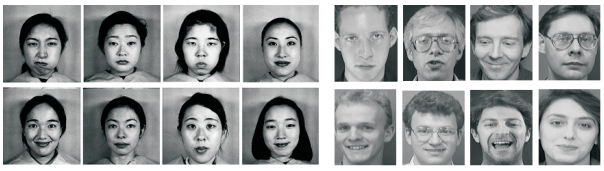
3) 观察表 5 可以发现, KT-MEC 的聚类性能始终不会差于经典的 MEC 算法, 这是因为 KT-MEC 的迁移学习程度可通过调节迁移参数 λ 和 η 进行控制, 该策略有效地抑制了负迁移现象的产生, 保证了 KT-MEC 算法较 MEC 算法更有效.

表 5　6 种算法在 NG 20 数据集上的聚类性能对比

数据集	评价指表	算　法					
		LSSMTC	combKM	MEC	STC	TSC	KT-MEC
rec VS talk	NMI-mean	0.081 8	0.057 2	0.269 1	0.186 5	0.422 4	**0.647 0**
	NMI-std	1.46e-17	0.020 1	0	0.005 5	0	1.17e-16
	RI-mean	0.502 1	0.500 2	0.596 0	0.574 7	0.735 9	**0.859 3**
	RI-std	0	0.000 4	1.17e-16	0.007 8	0	1.17e-16
comp VS sci	NMI-mean	0.019 6	0.002 1	0.104 9	0.124 0	**0.307 3**	0.142 2
	NMI-std	1.83e-18	0	0.071 7	0.002 7	0	0
	RI-mean	0.499 0	0.499 0	0.532 1	0.537 2	**0.678 1**	0.529 8
	RI-std	5.85e-17	5.85e-17	0.026 2	0.014 0	0	0

4.2.3　人脸识别实验和结果分析

实验使用 JAFFE 和 ORL 两个人脸数据库, 按人脸来源分为两个迁移场景: JAFFE 场景从 JAFFE 库中选择 8 个不同人, 每个人选择 17 幅图像, 15 幅作为源域数据集, 2 幅作为目标域数据集, 对每幅人脸进行旋转 10° 和 20° 形成最终的数据集, 如图 3(a) 所示; ORL 场景从 ORL 库中选择 8 个不同人的图像, 8 幅作为源域数据集, 2 幅作为目标域数据集, 对每幅人脸进行旋转 10° 和 20° 形成最终的数据集如图 3(b) 所示. 将图像像素灰度值作为每幅图像的原始特征, 并利用 PCA 方法对原始图像进行特征降维预处理. JAFFE 场景和 ORL 场景的数据集如表 6 所示. 6 个对比算法在两个迁移场景上的聚类结果如表 7 所示. 实验结果进一步表明了本文 KT-MEC 算法的有效性, 对于 NMI 指标而言, KT-MEC 相对于 MEC 的性能提升平均达到 70% 以上.



(a) JAFFE人脸数据集　　(b) ORL人脸数据集

图 3　人脸数据集

表 6　人脸数据集具体参数

数据类别	数据集	样本个数	维数	类别
JAFFE	源域数据	360	407	8
	目标域数据	48	407	
ORL	源域数据	192	239	8
	目标域数据	48	239	

表 7 6种算法人脸的聚类性能对比

数据集	评价指标	算法					
		LSSMTC	combKM	MEC	STC	TSC	KT-MEC
JAFFE	NMI-mean	0.458 4	0.264 4	0.309 1	0.385 4	0.282 2	**0.496 7**
	NMI-std	0	0.078 7	0.036 4	0.018 0	0.019 2	0
	RI-mean	0.670 2	0.440 2	0.801 7	0.820 2	0.804 5	**0.828 0**
	RI-std	1.17e-16	0.108 1	0.007 5	0.002 9	0.003 0	1.17e-16
ORL	NMI-mean	0.358 2	0.212 4	0.290 9	0.331 0	0.295 0	**0.521 8**
	NMI-std	0	0.095 4	0	0.018 3	0.005 4	0
	RI-mean	0.774 8	0.587 0	0.799 6	0.811 6	**0.812 4**	0.767 7
	RI-std	0	0.180 6	0	0.003 4	0.000 4	1.17e-16

5 结 论

本文针对传统划分模糊聚类方法在数据量不足或受干扰及噪声污染时聚类性能低下的问题展开研究, 首先提出了两种适用于模糊划分聚类的迁移学习机制, 即历史类中心迁移学习机制和关于历史类中心的隶属度迁移学习机制, 并使之与极大熵聚类问题相结合, 从而提出具有迁移学习能力的极大熵聚类技术, 即 KT-MEC 算法. 仿真实验表明, 所提出算法的聚类有效性明显增强, 且有效避免了负迁移现象的发生, 具有较好的源域数据隐私保护能力.

参考文献(References)

[1] Gao J, Fan W, Jiang J, et al. Knowledge transfer via multiple model local structure mapping[C]. Proc of The 14th Acm Sigkdd Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 283-291.

[2] Tao J, Chung F L, Wang S T. On minimum distribution discrepancy support vector machine for domain adaptation[J]. Pattern Recognition, 2012, 45(11): 3962-3984.

[3] Yang P, Tan Q, Ding Y, Bayesian task-level transfer learning for non-linear regression[C]. 2008 Int Conf on Computer Science and Software Engineering. Wuhan: IEEE, 2008: 62-65.

[4] Deng Z H, Jiang Y Z, Choi K S, et al, Knowledge-leverage-based TSK fuzzy system modeling[J]. IEEE Trans on Neural Netw Learning System, 2013, 24(8): 1200-1212.

[5] Wang Z, Song Y Q, Zhang C S. Transferred dimensionality reduction[C]. Lecture Notes in Computer Science. Berlin: Springer Heidelberg, 2008: 550-565.

[6] Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction[C]. The 23rd Conf on Artificial Intelligence. Illinois : AAAI, 2008, 8: 677-682.

[7] Dai W Y, Yang Q, Xue G R, et al. Self-taught clustering[C]. Proc of the 25th Int Conf on Machine Learning. Helsinki: ACM, 2008, 8: 200-207.

[8] Jiang W H, Chung F L. Transfer spectral clustering[C]. Machine Learning and Knowledge Discovery in Databases. Berlin: Springer Heidelberg, 2012: 789-803.

[9] Höppner F, Klawonn F, Kruse R. Fuzzy cluster analysis: Methods for classification, data analysis and image recognition[M]. New York: Wiley, 1999: 197-221.

[10] Bezdek J C, Hathaway R J. Numerical convergence and interpretation of the fuzzy c-shells clustering algorithms[J]. IEEE Trans on Neural Netw, 1992, 3(5): 787-793.

[11] Wu K L, Yu J, Yang M S. A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests[J]. Pattern Recognition Lett, 2005, 26(5): 639-652.

[12] Karayiannis N B. MECA: Maximum entropy clustering algorithm[C]. Proc of the 3rd IEEE Conf on IEEE World Congress on Computational Intelligence. Orlando: IEEE, 1994: 630-635.

[13] Li R P, Mukaidono M A. A maximum entropy approach to fuzzy clustering[C]. Proc on IEEE Int Conf Fuzzy System. Yokohama: IEEE, 1995: 2227-2232.

[14] Wang S T, Chung K, Deng Z H, et al. Robust maximum entropy clustering with its labeling for outliers[J]. Soft Compt, 2006, 10(7): 555-563.

[15] Zhi X B, Fan J L, Zhao F. Fuzzy linear discriminant analysis-guided maximum entropy fuzzy clustering algorithm[J]. Pattern Recognition, 2013, 46(6): 1604-1615.

[16] Qian P Q, Chung F L, Wang S T, et al. Fast graph-based relaxed clustering for large data sets using minimal enclosing ball[J]. IEEE Trans on Systems, Man and Cybernetics, 2012, 42(3): 672-687.

[17] Desgraupes B. Clustering indices[M]. Paris: University of Paris Ouest, 2013: 70-84.

[18] Gu Q, Zhou J. Learning the shared subspace for multi-task clustering and transductive transfer classification[C]. The 9th IEEE Int Conf on Data Mining. Miami: IEEE, 2009: 159-168.

[19] Liu J, Mohammed J, Carter J, et al. Distance-based clustering of CGH data[J]. Bioinformatics, 2006, 22(16): 1971-1978.

[20] Dai W Y, Xue G R, Yang Q, et al. Co-clustering based classification for out-of-domain documents[C]. The 13th ACM Sigkdd Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2007: 210-219.

[21] McCallum A K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering [DB/OL]. [1996-12-15]. (2014-01-15). http://www.cs.cmu.edu/mccallum/bow.

(责任编辑：郑晓蕾)