DOI:10.11992/tis.201603005
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.tp.20160824.0928.004.html

知识迁移的极大熵聚类算法及其在纹理图像分割中的应用

程旸,蒋亦樟,钱鹏江,王士同
(江南大学 数字媒体学院,江苏 无锡 214122)

摘　要:本文研究了一种新型的基于知识迁移的极大熵聚类技术。拟解决两大挑战性问题:1)如何从源域中选择合适的知识对目标域进行迁移学习以最终强化目标域的聚类性能;2)若存在源域聚类数与目标域聚类数不一致的情况时,该如何进行迁移聚类。为此提出一种全新的迁移聚类机制,即基于聚类中心的中心匹配迁移机制。进一步将该机制与经典极大熵聚类算法相融合提出了基于知识迁移的极大熵聚类算法(KT-MEC)。实验表明,在不同迁移场景下的纹理图像分割应用中,KT-MEC 算法较很多现有聚类算法具有更高的精确度和抗噪性。

关键词:迁移学习;中心迁移匹配;极大熵聚类;纹理图像分割;抗噪性
中图分类号: TP181　**文献标志码**:A　**文章编号**:1673-4785(2017)02-0179-09

中文引用格式:程旸,蒋亦樟,钱鹏江,等. 知识迁移的极大熵聚类算法及其在纹理图像分割中的应用[J]. 智能系统学报, 2017, 12(2): 179-187.
英文引用格式:CHENG Yang, JIANG Yizhang, QIAN Pengjiang, et al. A maximum entropy clustering algorithm based on knowledge transfer and its application to texture image segmentation[J]. CAAI transactions on intelligent systems, 2017, 12(2): 178-187.

A maximum entropy clustering algorithm based on knowledge transfer and its application to texture image segmentation

CHENG Yang, JIANG Yizhang, QIAN Pengjiang, WANG Shitong
(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract: In this paper, we propose a novel technique for maximum entropy clustering (MEC) based on knowledge transfer. More specifically, we aim to solve the following two challenging questions. First, how can knowledge be appropriately selected from a source domain to enhance clustering performance in the target domain via transfer learning? Second, how best do we conduct transfer clustering if the number of clusters in the source domain and the target domain are inconsistent? To address these questions, we designed a new transfer clustering mechanism called the central matching transfer mechanism, which we based on clustering centers. Further, we developed a knowledge-transfer-based maximum entropy clustering (KT-MEC) algorithm by incorporating our mechanism into the classic MEC approach. Our experimental results reveal that our proposed KT-MEC algorithm achieves a higher level of accuracy and better noise immunity than many existing methods when applied to texture image segmentation in different transfer scenarios.

Keywords: transfer learning; center transfer matching; maximum entropy clustering; texture image segmentation; robustness

在实际生产中,大部分机器学习方法处理的对象均为含噪数据集且存在数据量不足的问题。如对于图像分割[1]任务而言,图像数据往往含有很大的噪声。图像数据含噪程度越高,使用的机器学习方法对其进行分割的性能就变得越弱。一般来说,无

收稿日期:2016-03-04.　**网络出版日期**:2016-08-24.
基金项目:国家自然科学基金项目(61572236);江苏省自然科学基金项目(BK20160187);江苏省产学研前瞻性联合研究项目(BY2013015-02).
通信作者:蒋亦樟. E-mail:jyz0512@163.com.

监督的聚类方法通常用来获得图像的分割结果[2-3]，比较著名的算法有模糊 C 均值算法(FCM)[4]、可能性聚类算法(PCM)[5]、极大熵聚类算法[6]等。这些方法虽简单实用，但其对于含噪图像数据的分割效果并不理想。尽管已有学者致力于解决该问题，但效果并不明显。

1 问题描述

迁移学习技术[7]的提出，为我们提供了一种新的解决问题的思路。传统的机器学习假设训练数据与测试数据服从相同的数据分布。然而，大量实际情况中并不满足这种同分布假设。从另外一个角度上看，如果我们已经有了大量的、在不同分布下的训练数据，完全丢弃这些数据是非常浪费的。如何合理地利用这些数据就是迁移学习要解决的问题。迁移学习可以从现有的数据中迁移知识，用来帮助将来的学习。迁移学习的目标是将从一个环境中学到的知识用来帮助新环境中的学习任务，其学习过程类似人类的学习和思维方式。我们面临的问题如图 1 所示。

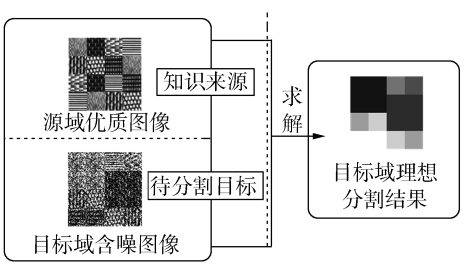


图 1 问题描述

Fig.1 The description of the problem

源域的数据中往往存在一部分数据为可用数据，如源域优质图像，目标域的数据通常呈现数据不足或噪声污染严重[8-9]等情况，如目标域含噪图像。如何才能得到最接近目标域理想分割的效果图，如果能够将源域的知识成功迁移到目标域中进行学习，是否能够大幅提高图像分割性能呢？

为了验证本文的设想，实现提高图像分割性能的目标，本文将迁移学习方法融入到经典的极大熵聚类算法[10](maximum entropy clustering，MEC)中，以提高极大熵算法的聚类性能，进而提高该算法对图像分割的性能。在将迁移学习策略融入到极大熵聚类算法的过程中，我们面临的挑战有：1)选择源域的何种知识进行迁移学习以增强目标域的聚类性能；2)当源域和目标域的聚类数不一致时如何迁移。

针对挑战 1)，本文选用聚类中心作为迁移知识，因源域的聚类中心是各类所包含点的高度浓缩，亦是各类的代表点，将其作为聚类中的高级知识具有更强的指导性；针对挑战 2)，本文提出了一个中心迁移匹配机制用于处理源域和目标域聚类数不一致的情况。无论源域与目标域的聚类数是否相同，该中心迁移匹配机制均可适用，且能够找到源域与目标域类中心的最佳匹配关系。将上述迁移知识与中心迁移匹配机制融入到经典的极大熵聚类算法中，本文提出了一种全新的基于知识迁移的极大熵聚类算法，并将该算法成功应用于纹理图像分割中。实验结果表明，本文所提出的基于知识迁移的极大熵聚类算法在不同的迁移场景下对于纹理图像的分割性能均优于其他迁移以及非迁移聚类算法。本文工作的创新主要涵盖以下几点：

1)确定了源域中哪种知识能够进行有效迁移，即从源域数据中获取的聚类中心知识可以用来指导并增强目标域的聚类性能；

2)给出了一种解决源域与目标域聚类数不同时，如何进行有效迁移的途径，即提出了一种通用的中心迁移匹配机制，不仅能够有效解决源域与目标域聚类数不相同时的迁移问题，还能指导源域、目标域聚类数相同时，各类中心如何一一对应的问题。

3)将上述两个问题的解决策略融入到极大熵聚类算法后，本文提出了一种新的基于知识迁移的极大熵聚类算法，实验表明该算法的聚类性能较其他迁移聚类算法以及非迁移聚类算法在处理不同迁移场景下的纹理分割图像时，具有更加优良的性能。

本文所用的符号说明如表 1 所示。

表 1 符号说明

Table 1 The explanation of some notations

符号	描述
$\boldsymbol{U}=[u_{ij}]_{C\times N}$	隶属度矩阵，U_{ij} 代表第 i 个数据属于第 j 个聚类中心的可能性
$\boldsymbol{P}=[p_{jk}]_{c_t\times c_s}$	知识迁移隶属度矩阵，p_{jk} 代表目标域第 j 个中心属于源域第 k 个聚类中心的程度
$\boldsymbol{V}=[\boldsymbol{V}_1 \quad \cdots \quad \boldsymbol{V}_C]^{\mathrm{T}}$ $\boldsymbol{V}_i=[v_{i1} \quad \cdots \quad v_{iD}]^{\mathrm{T}}$	聚类中心矩阵 $\boldsymbol{V}_i$ 代表第 i 个聚类中心
γ	熵的正则化参数
λ	迁移平衡参数
C	聚类数
T	总的迭代次数
N	样本总数
D	特征总数
s	如果将 s 作为一个符号的下标，表示这个符号属于源域
t	如果将 t 作为一个符号的下标，表示这个符号属于目标域

2　相关工作

2.1　经典 MEC 算法

MEC 聚类算法是基于划分的聚类算法中最具代表性的算法之一,该算法的数学表达式简单明了、物理意义明确,是广大学者较常使用的聚类算法,关于 MEC 算法的变形算法较经典的如文献[10]。特别是在针对含有噪声的纹理图像的分割中,MEC 聚类算法相比经典的模糊 C 均值聚类 FCM 以及可能性聚类 PCM 等聚类算法具有更好的抗噪性,进而能够获得更好的聚类性,使得分割结果更加逼近理想分割结果。综上,本文选用了 MEC 算法。MEC 算法的函数表达式为

$$\min_{\boldsymbol{U},\boldsymbol{V}}\Big(\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\|\boldsymbol{x}_j-\boldsymbol{V}_i\|^2+\gamma\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}\ln\mu_{ij}\Big)$$
$$\text{s.t.}\quad 0\leqslant\mu_{ij}\leqslant 1,\ \sum_{i=1}^{C}\mu_{ij}=1,$$
$$1\leqslant i\leqslant C,1\leqslant j\leqslant N \tag{1}$$

式中:$\boldsymbol{x}_j$ 为第 j 个数据样本,$\boldsymbol{V}_i$ 为第 i 个聚类中心,u_{ij} 为样本 $\boldsymbol{x}_j$ 属于聚类中心 $\boldsymbol{V}_i$ 的隶属度,C 为聚类数,N 为样本总数,γ 为熵的正则化参数,$\|\boldsymbol{x}_j-\boldsymbol{V}_i\|^2$ 代表样本 $\boldsymbol{x}_j$ 与聚类中心 $\boldsymbol{V}_i$ 之间的距离。

由拉格朗日乘子法则,求解式(1),解得聚类中心 $\boldsymbol{V}_i$ 和隶属度 $\boldsymbol{U}_{ij}$ 的表达式为

$$\boldsymbol{V}_i=\frac{\sum_{j=1}^{N}\mu_{ij}\boldsymbol{x}_j}{\sum_{j=1}^{N}\mu_{ij}},i=1,2,\cdots,C \tag{2}$$

$$\mu_{ij}=\frac{\exp\left(-\dfrac{\|\boldsymbol{x}_j-\boldsymbol{V}_i\|^2}{\gamma}\right)}{\sum_{k=1}^{C}\exp\left(-\dfrac{\|\boldsymbol{x}_j-\boldsymbol{V}_k\|^2}{\gamma}\right)} \tag{3}$$

$i=1,2,\cdots,C;\ j=1,2,\cdots,N$。

MEC 算法步骤如下:

1)给定聚类数 C,样本总数 N,正则化参数 γ,聚类精度 ε,最大迭代次数 T,初始化隶属度矩阵 $\boldsymbol{U}$ 和聚类中心 $\boldsymbol{V}$;

2)根据式(2)更新聚类中心矩阵 $\boldsymbol{V}$;

3)根据式(3)更新隶属度矩阵 $\boldsymbol{U}$;

4)当 $\|\boldsymbol{U}(t+1)-\boldsymbol{U}(t)\|<\varepsilon$ 或迭代次数达到最大迭代次数 T 时,算法运行终止,否则,返回 2);

5)算法收敛后,输出聚类中心 $\boldsymbol{V}$ 和隶属度矩阵 $\boldsymbol{U}$。

2.2　相关迁移聚类算法

近年来,迁移聚类算法及其相关算法的研究已受到许多专家学者的关注,本文将研究中较有价值的文献罗列如下:文献[11]提出了一种自学聚类算法,该算法是第 1 个基于互信息的迁移聚类算法,但是由于该算法运行的前提是假定源域数据是可用的,这在实际生产应用中并不切实际,所以该算法具有一定的局限性;文献[12]提出了一种基于谱聚类的迁移聚类算法,该算法主要针对光谱聚类;文献[13]提出了一种极大熵的迁移聚类算法,该算法提出了基于类中心和隶属度的两种知识迁移机制,但该算法并未解决当源域目标域聚类数不一致时,如何进行迁移的问题。除了直接提出的迁移聚类算法,还存在如协同聚类[14]、多任务聚类[15]、联合聚类[16]、半监督聚类[17]等具有相关性的聚类算法。其中,协同聚类算法的核心思想为结合样本间不同的协作能力形成拉动效应,共同推动事物的发展,从而提高样本的整体聚类精度。多任务聚类的核心思想为多个聚类任务同时进行,各个聚类任务之间相互协调配合,以提高聚类性能。联合聚类顾名思义就是联合多个聚类算法进行一定关系的联合使用,聚类精度的提高对于具体聚类算法的选择比较敏感。半监督聚类算法需要已知一部分数据样本的标签,根据这些标签来指导整个样本数据的聚类过程,从而提高聚类性能。

现有的迁移聚类算法及其相关算法在处理含噪的图像分割数据时,均存在各种问题。如文献[13]提出的迁移聚类算法无法解决当源域与目标域的图像分割数不一致时,如何实现迁移的问题。对于其他相关算法如联合算法来说,图像本身还有噪声,经过层层的聚类算法进行处理,误差被层层放大,最终的聚类性能则被削弱。本文所做研究主要针对纹理图像分割进行展开,我们将在下一节针对算法的抗噪性、源域目标域聚类数是否一致等问题进行详细描述。

3　基于知识迁移的 MEC 聚类算法

3.1　基于聚类中心的知识迁移机制

源域中存在许多知识可用于迁移到目标域中进行学习。问题在于在具体选择时,应该选择哪种或哪几种知识的组合进行迁移。源域中存在可以迁移的知识主要有:聚类中心、隶属度、数据样本以及其他经过二次或多次处理后获得的知识。考虑到源域的聚类中心具有较高的数据集中特征,且该知识作为自然聚类知识的核心,本文最终选择了聚类中心作为知识迁移的对象。基于中心迁移的表达式计算的是源域的聚类中心 $\boldsymbol{V}_s$ 与目标域 $\boldsymbol{V}_t$ 之间距离和。

$$\Delta_1(\boldsymbol{V}_s,\boldsymbol{V}_t)=\lambda\sum_{j=1}^{C_t}\|\boldsymbol{V}_{j,t}-\boldsymbol{V}_{j,s}\|^2 \tag{4}$$

式中：λ 为迁移平衡参数，一般大于 0，其值越大，表示源域知识在目标域中所占分量越大；C_t 为目标域聚类数；$\boldsymbol{V}_{j,t}$ 为目标域中第 j 个聚类中心；$\boldsymbol{V}_{j,s}$ 为源域中第 j 个聚类中心。

3.2 基于聚类中心的迁移匹配机制

式(4)尽管实现了源域知识向目标域迁移进行指导学习的目的，但其并未解决源域与目标域的聚类数不相同时，如何进行迁移和中心间的匹配问题。本小节，我们将致力于探讨能否确定一个通用的准则，无论源域与目标域的聚类数是否一致均能自适应地匹配。为了解决上述问题，本文引入了模糊聚类理论来解决该问题，从而提出了一种中心迁移匹配机制。中心迁移匹配机制的表达式为

$$\begin{aligned}&\Delta_2(\boldsymbol{P}_{t,s},\boldsymbol{V}_t,\boldsymbol{V}_s)=\lambda\sum_{j=1}^{C_t}\sum_{k=1}^{C_s}p_{jk}\parallel\boldsymbol{V}_{j,t}-\boldsymbol{V}_{k,s}\parallel^2\\&\text{s.t. }p_{jk}\in[0,1],\sum_{k=1}^{C_s}p_{jk}=1,1\leqslant i\leqslant N_t\\&1\leqslant j\leqslant C_t,1\leqslant k\leqslant C_s\end{aligned}\tag{5}$$

式(5)解决了源域的聚类中心 $\boldsymbol{V}_s$ 与目标域 $\boldsymbol{V}_t$ 之间的匹配问题。其中，参数 $\boldsymbol{P}_{t,s}$ 为知识迁移隶属度，p_{jk} 表示目标域的第 j 个类中心与源域的第 k 个类中心进行匹配的隶属度。当 $p_{jk}\to1$，表示目标域的第 j 个类中心完全匹配源域的第 k 个类中心；当 $p_{jk}\to0$，表示目标域的第 j 个类中心不匹配源域的第 k 个类中心，若出现不匹配的情况，源域中未找到匹配聚类中心的那个聚类中心将会从源域的聚类中心中删除掉。N_t 为目标域数据样本的大小，C_t 为目标域聚类数，C_s 为源域聚类数。

3.3 基于知识迁移的极大熵聚类算法

将上述知识迁移机制与知识匹配机制融入到 MEC 聚类算法后，本文提出一种基于知识迁移的极大熵聚类算法。该算法的流程主要分为两个阶段，流程图如图 2 所示。

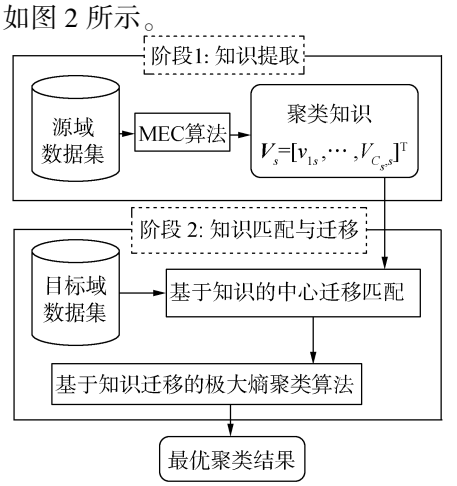


图 2 KT-MEC 算法流程图

Fig.2 The flowchart of KT-MEC algorithm

1) 知识提取

利用经典的极大熵聚类算法对源域的数据集进行聚类，得到源域的聚类中心 $\boldsymbol{V}_s$。知识提取的表达式为

$$\begin{aligned}&J_{\text{MEC}}=\sum_{i=1}^{C_s}\sum_{j=1}^{N_s}\mu_{ij,s}\parallel\boldsymbol{x}_{j,t}-\boldsymbol{V}_{i,s}\parallel^2+\gamma\sum_{i=1}^{C_s}\sum_{j=1}^{N_s}\mu_{ij,s}\ln\mu_{ij,s}\\&\text{s.t. }\mu_{ij,s}\in[0,1],\sum_{i=1}^{C_s}\mu_{ij,s}=1,1\leqslant i\leqslant C_s,1\leqslant j\leqslant N_s\end{aligned}\tag{6}$$

通过求解式(6)，得到源域聚类中心 $\boldsymbol{V}_s$。

2) 知识匹配与迁移

利用中心迁移匹配机制将阶段 1 得到的聚类知识进行自适应匹配，使源域中的聚类中心(知识)能够与目标域中的聚类中心进行完美匹配，以解决源域和目标域不同类时的迁移问题。同时，将匹配后的源域知识迁移到目标域中加以利用。结合极大熵聚类算法，基于知识迁移的极大熵聚类算法(KT-MEC)，该算法的函数表达式为

$$\begin{aligned}&J_{\text{KT-MEC}}=\sum_{i=1}^{C_t}\sum_{j=1}^{N_t}\mu_{ij,t}\parallel\boldsymbol{x}_{j,t}-\boldsymbol{V}_{i,t}\parallel^2+\\&\gamma\sum_{i=1}^{C_t}\sum_{j=1}^{N_t}\mu_{ij,t}\ln\mu_{ij,t}+\lambda\sum_{i=1}^{C_t}\sum_{k=1}^{C_s}p_{ik}\parallel\boldsymbol{V}_{i,t}-\boldsymbol{V}_{k,s}\parallel^2+\\&\eta\sum_{i=1}^{C_t}\sum_{k=1}^{C_s}p_{ik}\ln p_{ik}\\&\text{s.t.}u_{ij,t}\in[0,1],\sum_{i=1}^{C_t}\mu_{ij,t}=1,p_{ik}\in[0,1],\sum_{k=1}^{C_s}p_{ik}=1\\&1\leqslant j\leqslant N_t,1\leqslant i\leqslant C_t,1\leqslant k\leqslant C_s\end{aligned}\tag{7}$$

式中：$u_{ij,t}$ 为目标域隶属度，$x_{j,t}$ 为目标域第 j 个样本数据，$\boldsymbol{V}_{i,t}$ 为目标域第 i 个聚类中心，γ 为熵的正则化参数，C_t 为目标域聚类数，N_t 为目标域样本总数，λ 为知识迁移的平衡系数，p_{ik} 表示目标域的第 i 个类中心迁移到源域的第 k 个类中心的知识迁移隶属度，$\boldsymbol{V}_{k,s}$ 为源域的第 k 个类中心，η 为迁移项的正则化参数。通过拉格朗日乘子法最小化式(7)，各参数表达式如下：

目标域隶属度 $u_{ij,t}$：

$$\mu_{ij,t}=\frac{\exp\left(-\dfrac{\parallel\boldsymbol{x}_{j,t}-\boldsymbol{V}_{i,t}\parallel^2}{\gamma}\right)}{\sum_{l=1}^{C_t}\exp\left(-\dfrac{\parallel\boldsymbol{x}_{j,t}-\boldsymbol{V}_{l,t}\parallel^2}{\gamma}\right)}$$

目标域聚类中心 $v_{i,t}$：

$$\boldsymbol{v}_{i,t}=\frac{\sum_{j=1}^{N_t}\mu_{ij,t}x_{j,t}+\lambda\sum_{k=1}^{C_s}p_{ik}\boldsymbol{V}_{k,s}}{\sum_{j=1}^{N_t}\mu_{ij,t}+\lambda\sum_{k=1}^{C_s}p_{ik}}$$

知识迁移隶属度 p_{ik} :

$$p_{ik}=\frac{\exp\left(-\frac{\lambda \parallel \boldsymbol{V}_{i,t}-\boldsymbol{V}_{k,s}\parallel^2}{\eta}\right)}{\sum_{l'=1}^{c_s}\exp\left(-\frac{\lambda \parallel \boldsymbol{V}_{i,t}-\boldsymbol{V}_{l',s}\parallel^2}{\eta}\right)}$$

通过上述两个阶段的流程,将各源域与目标域的相关数据带入到各表达式中,得到最终的聚类结果。KT-MEC 聚类算法的详细步骤如下:

输入　源域数据集 x_s,目标域数据集 x_t,源域聚类数 C_s,目标域聚类数 C_t,熵的正则化参数 γ,收敛精度 ε,最大迭代次数 T;

输出　目标域隶属度 $\boldsymbol{U}_t$,目标域聚类中心 $\boldsymbol{V}_t$。

知识提取阶段:

1)随机初始化源域的隶属度矩阵 $\boldsymbol{U}_s$;

2)利用式(2)求得源域的聚类中心 $\boldsymbol{V}_s$;

3)利用式(3)求得源域的隶属度 $\boldsymbol{U}_s$;

4)满足迭代终止条件则输出源域聚类中心 $\boldsymbol{V}_s$ 并终止算法,否则返回 2)。

知识匹配与迁移阶段:

1)随机初始化目标域的隶属度矩阵 $\boldsymbol{U}_t$ 以及聚类中心 $\boldsymbol{V}_t$;

2)利用式(8)求得目标域的隶属度矩阵 $\boldsymbol{U}_t$;

3)利用式(9)求得目标域聚类中心矩阵 $\boldsymbol{V}_t$;

4)利用式(10)求得目标域的知识迁移隶属度矩阵 $\boldsymbol{P}_{ts}$;

5)如满足迭代终止条件则输出目标域隶属度矩阵 $\boldsymbol{U}_t$,聚类中心 $\boldsymbol{V}_t$,并终止算法,否则返回 2)。

4　实验与分析

为了评估本文所提 KT-MEC 聚类算法的性能,实验所使用的对比算法有:非迁移 MEC 聚类算法、自学聚类算法(STC)[11]、迁移谱聚类算法(TSC)[12]、DRCC 协同聚类算法[15]、CombKM 多任务聚类算法[15]。本文实验所用数据集为 Brodatz 纹理图像分割[17]数据集。

Brodatz 纹理图像由 7 个基本纹理图像(D_3、D_6、D_{21}、D_{49}、D_{53}、D_{56}、D_{93})合成,具体见图 3。合成纹理图像的大小被重新调整为 100 像素×100 像素。为了模拟真实数据集环境,本文将不同标准偏差的高斯噪声添加到各个纹理图像中。实验中,图 3(a)为源域的图像数据,图 3(b)~(i)为在不同的目标域中的图像数据。为了模拟不同的迁移场景,我们设计了两种不同迁纹理图像分割任务,目标域图像 T_1 ~ T_4 与源域图像的类别数均为 7,σ =0.1,0.2,0.0,0.1;目标域图像 T_5 ~ T_8 与源域图像的类别数分别为 3、4、5、6,σ =0.1。

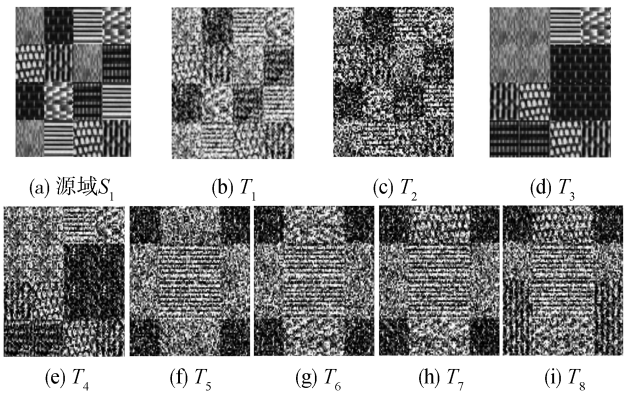


图 3　源域及不同情况下目标域的纹理图像数据

Fig.3　Texture image datasets of one source domain and some different target domains

理想分割图可用来为各算法的分割性能优劣作参考,理想的纹理分割结果如图 4 所示。

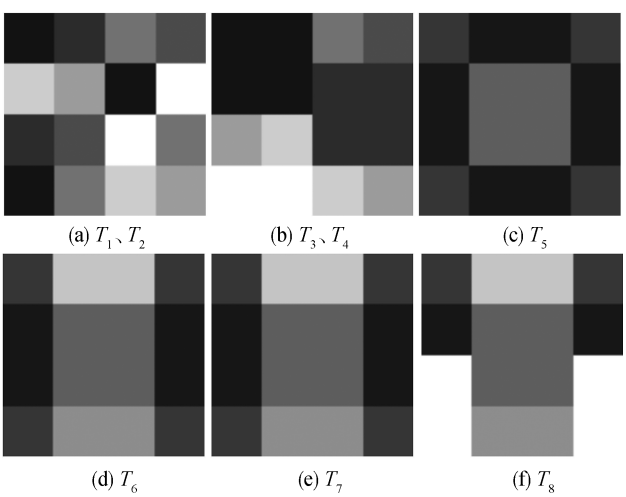


图 4　不同纹理图像的理想分割结果

Fig.4　Ideal segmentation result of different texture images

纹理图像分割的过程概括如下。文献[19]先采用 Gabor 滤波器在 6 个方向提取纹理图像特征的滤波器组。每个纹理图像的数据集包含 30 维特征,数据集大小为 10 000。不同算法得到的类被认为分割图像的一个区域。

4.1　实验参数设置

通常用来衡量聚类算法性能的指标有:NMI、RI、Entropy、F-measure 等,本文主要采用以下两种评估指标:

$$\mathrm{NMI}=\frac{\sum_{i=1}^{C}\sum_{j=1}^{C}(N_{i,j}\log N\cdot N_{i,j})/N_i\cdot N_j}{\sqrt{\sum_{i=1}^{C}N_i\log N_i/N\cdot\sum_{j=1}^{C}N_j\log N_j/N}}$$

$$\mathrm{RI}=\frac{f_{00}+f_{11}}{N(N-1)/2}$$

式中：$N_{i,j}$ 表示第 i 个聚类与类 j 的契合程度，N_i 表示第 i 个聚类所包含的数据样本量，N_j 表示类 j 所包含的数据样本量，而 N 表示整个数据样本的总量大小。RI 表达式中的 f_{00} 表示数据点具有不同的类标签并且属于不同类的配对点数目，f_{11} 则表示数据点具有相同的类标签并且属于同一类的配对点数目，而 N 表示整个数据样本的总量大小。NMI、RI 两种评价指标的取值范围均为[0,1]，取值越大表明算法的性能越好。

在本文所使用的迁移算法中，KT-MEC 算法的熵正则化参数 $\gamma \in \{0:0.05:1\}$，迁移平衡因子 $\lambda \in \{0.1,0.5,1,5,10,50,100,500,1\,000\}$，迁移隶属度的正则化参数 $\eta \in \{0:0.05:1\}$。TSC 算法和 STC 算法的参数设置详见文献[11]和文献[12]。

在本文所使用的对比算法中，极大熵聚类 MEC 的熵正则化参数 $\gamma \in \{0:0.05:1\}$，协同聚类 DRCC 的正则化参数 λ 和 μ 取值为 $\{0.1,1,10,100,500,1\,000\}$。

上述所有参数均由网格搜索[18]得到最优值，实验结果均为运行算法 15 次的结果取均值及方差所得。实验均在 MARTLAB 8.1.0.604 (R2013a)平台下完成，操作系统为 64 位 Windows7，CPU 为 Intel(R) Core(TM) i3-3240 3.40 GHz，内存为 4 GB。

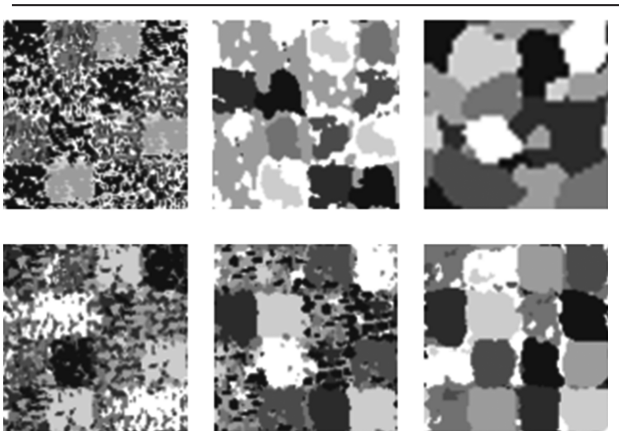
4.2 聚类数相同的纹理图像分割

表 2 与图 5 分别为源域与目标域聚类数相同时，各算法对纹理图像进行分割时的聚类性能对比与图像分割结果对比。

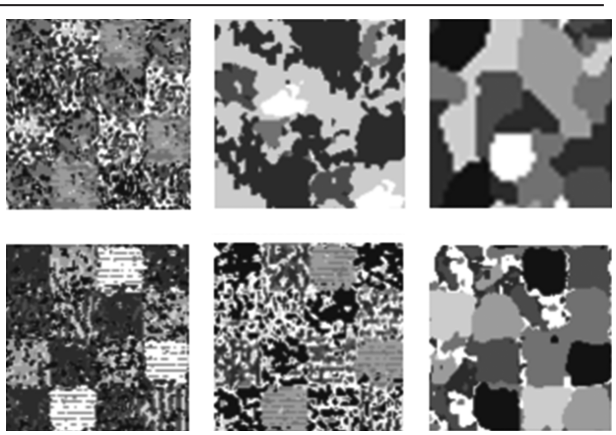
表 2 源域与目标域的聚类数相同时的各算法聚类性能对比

Table 2 Performance comparison of algorithms when the number of clusters of source domain and target domain are same

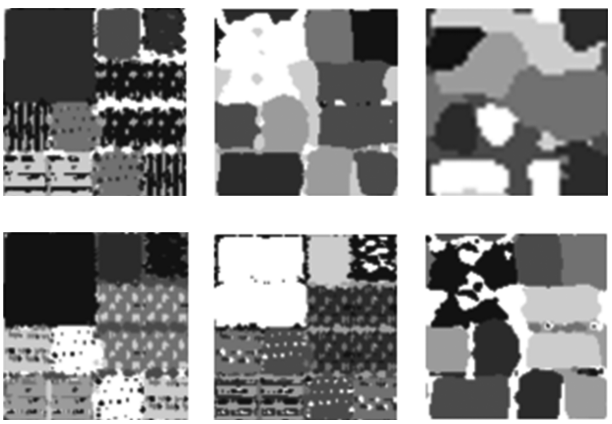
数据集	评价指标	MEC	CombKM	DRCC	STC	TSC	KT-MEC
T_1	NMI-mean	0.415 1	0.250 0	0.248 0	0.498 6	0.513 3	0.633 6
	NMI-std	0.005 2	0.030 8	0.021 3	0	0.006 6	0.004 9
	RI-mean	0.826 8	0.753 2	0.790 5	0.869 0	0.877 2	0.906 3
	RI-std	0.008 7	0.015 4	0.003 7	0	0.002 2	0.001 1
T_2	NMI-mean	0.302 7	0.231 1	0.226 4	0.369 6	0.347 0	0.535 0
	NMI-std	0.004 8	0.048 1	0.018 8	0	0	0.009 1
	RI-mean	0.777 7	0.706 3	0.778 3	0.783 9	0.770 8	0.856 9
	RI-std	0.005 2	0.028 7	0.001 5	0	0	0.012 4
T_3	NMI-mean	0.603 9	0.609 2	0.342 2	0.651 1	0.610 4	0.619 8
	NMI-std	0.035 9	0.024 0	0.024 1	0	5.77×10^{-4}	0.003 2
	RI-mean	0.855 3	0.861 1	0.784 9	0.887 7	0.872 6	0.864 4
	RI-std	0.018 5	0.026 8	0.020 4	0	2.31×10^{-4}	0.000 8
T_4	NMI-mean	0.455 7	0.426 1	0.241 3	0.549 7	0.551 1	0.614 7
	NMI-std	0.019 7	0.019 3	0.014 3	0	0	0.001 1
	RI-mean	0.817 8	0.808 2	0.784 8	0.847 2	0.849 6	0.875 7
	RI-std	0.004 4	0.007 3	0.001 3	0	0	0.000 8



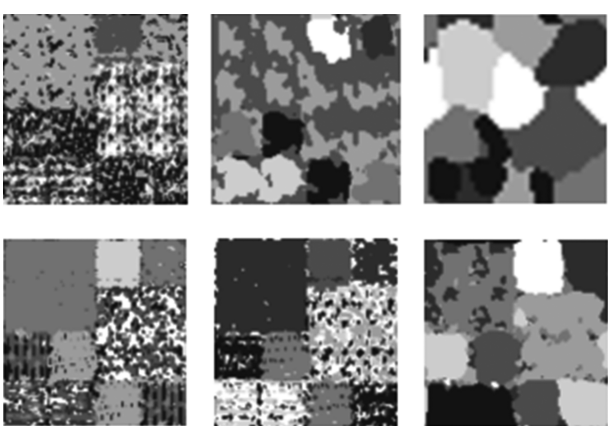
(a) 6 种算法分别在数据集 T_1 上的图像分割结果



(b) 6 种算法分别在数据集 T_2 上的图像分割结果



(c) 6 种算法分别在数据集 T_3 上的图像分割结果



(d) 6 种算法分别在数据集 T_4 上的图像分割结果

图 5　源域与目标域聚类数相同的含噪纹理图像分割结果

Fig.5　Segmentation results of clustering algorithms for noisy texture images with the same number of clusters between source domain and target domain

从表 2 和图 5 的聚类结果可以观察到,迁移聚类算法(STC、TSC、KT-MEC)在 $T_1 \sim T_4$ 数据集上取得了比传统的非迁移聚类算法更高的聚类精度。表 2 中 NMI 和 RI 值以及图 5 中可视化的分割结果,均表明本文所提出的 KT-MEC 聚类算法优于经典的 MEC 算法。以上结果进一步表明,在含噪的数据环境中,本文 KT-MEC 算法具有比 MEC 更好的鲁棒性,也进一步表明迁移学习技术是提高算法鲁棒性的有效途径。

如表 2 和图 5 的聚类结果所示,本文提出的 KT-MEC 算法与协同算法 DRCC 以及多任务聚类算法 CombKM 相比,本文算法仍然较优,这是因为多任务聚类与迁移聚类的原理明显不同。协同聚类与多任务聚类在集中完成多个聚类任务的同时,通过使用每个聚类任务的独立信息和多个聚类任务间的潜在相关信息,以获得良好的聚类性能。然而,在迁移聚类场景中,目标域的数据不能提供正确的聚类信息,这就会使得协同聚类和多任务聚类算法的聚类性能变弱。

此外,由于本文提出的 KT-MEC 算法较其他迁移聚类算法、协同聚类算法、多任务聚类算法具有更好的聚类性能,这进一步表明先进的集群知识(如聚类中心)可以被看作是一种有效的迁移知识,以提高目标域的聚类性能。这也表明本文提出的聚类中心自适应匹配机制能使源域的类中心与目标域的类中心进行成功匹配,达到知识迁移的目的。

4.3　聚类数不同的纹理图像分割

表 3 与图 6 分别为源域与目标域聚类数不同时,各算法对纹理图像进行分割时的聚类性能对比与图像分割结果对比。

由于协同聚类算法 DRCC、迁移聚类算法 STC 和 TSC 的聚类机制需要源域与目标域有相同的聚类数,所以这 3 种聚类算法不能在源域与目标域聚类数不同的迁移场景下运行。

表 3 和图 6 的实验结果表明本文提出 KT-MEC 聚类算法在图像分割性能上较经典的非迁移 MEC 算法以及 CombKM 算法具有更优的聚类性能。此外,得益于本文提出的基于知识的中心迁移机制,源域与目标域聚类数不同的迁移场景中的聚类结果表明了本文提出的基于知识的中心匹配机制可挖掘出源域和目标域之间完美的聚类中心的配对关系,进而确保知识迁移的质量。

表 3　源域与目标域的聚类数不同时的各算法聚类性能对比

Table 3　Performance comparison of algorithms when the number of clusters of source domain and target domain are different

数据集	评价指标	MEC	CombKM	DRCC	STC	TSC	KT-MEC
T_5	NMI-mean	0.464 4	0.555 7	—	—	—	0.650 1
	NMI-std	0.000 2	0.020 1	—	—	—	7.064×10^{-5}
	RI-mean	0.781 7	0.731 0	—	—	—	0.836 0
	RI-std	0.000 1	0.022 0	—	—	—	5.90×10^{-5}
T_6	NMI-mean	0.268 0	0.508 7	—	—	—	0.762 8
	NMI-std	0.005 0	0.069 0	—	—	—	1.11×10^{-16}
	RI-mean	0.687 2	0.657 8	—	—	—	0.916 8
	RI-std	0.002 0	0.088 8	—	—	—	1.35×10^{-16}

续表 3

数据集	评价指标	MEC	CombKM	DRCC	STC	TSC	KT-MEC
T_7	NMI-mean	0.291 0	0.576 9	—	—	—	0.7278
	NMI-std	0.008 0	0.018 9	—	—	—	0
	RI-mean	0.732 5	0.734 7	—	—	—	0.905 4
	RI-std	0.003 3	0.047 6	—	—	—	0
T_8	NMI-mean	0.203 8	0.572 8	—	—	—	0.691 4
	NMI-std	0.022 5	0.032 9	—	—	—	1.11×10^{-16}
	RI-mean	0.739 9	0.794 1	—	—	—	0.903 2
	RI-std	0.005 9	0.016 0	—	—	—	0

(a) 6 种算法分别在数据集 T_5 上的图像分割结果

(b) 6 种算法分别在数据集 T_6 上的图像分割结果

(c) 6 种算法分别在数据集 T_7 上的图像分割结果

(d) 6 种算法分别在数据集 T_8 上的图像分割结果

图 6 源域与目标域聚类数不同的含噪纹理图像分割结果

Fig.6 Segmentation results of clustering algorithms for noisy texture images with the different number of clusters between source domain and target domain

上述实验结果表明本文提出的 KT-MEC 聚类算法在不同的迁移场景中的聚类性能均优于现有的相关聚类算法。特别是,KT-MEC 聚类算法适用于一般的迁移场景,即无论是源域和目标域的聚类的数目是相同或不同时,本文 KT-MEC 算法均能适用且能获得比其他聚类算法更好的聚类结果。

5 结束语

本文研究是基于迁移学习的聚类算法,实验部分主要针对纹理图像的分割。本文算法对迁移聚类算法的贡献主要有两方面:1)确定了聚类中心作为迁移知识,实验证明了将聚类中心作为迁移知识能够更好地增强目标域的聚类性能;2)找到了一个解决无论源域与目标域的聚类数是否一致,都能够成功进行迁移的通用策略。基于上述工作,结合传统的非迁移极大熵聚类算法,本文提出了基于知识迁移的极大熵聚类算法,并将该算法与其他迁移算法、非迁移算法、协同聚类算法、多任务聚类算法等一系列相关算法进行了性能对比,实验表明本文 KT-MEC 聚类算法的性能在纹理图像分割上较其他算法具有更加优良的性能。KT-MEC 聚类算法不仅能够提高算法的聚类精度,增强图像的分割效果,还能适应不同迁移场景下的聚类任务,具有较强的鲁棒性。

虽然本文 KT-MEC 聚类算法在纹理图像的分割上具有较好的性能,但该算法的适应性上还需进行进一步的研究。随着数据的爆炸式增长,数据复杂性的迅速增加,KT-MEC 聚类算法是否能够适用于高维复杂数据还有待研究。

参考文献:

[1] ZHU Lin, CHUNG F L, WANG Shitong. Generalized fuzzy c-means clustering algorithm with improved fuzzy partitions [J]. IEEE transactions on systems, man, and cybernetics, part B (cybernetics), 2009, 39(3): 578-591.

[2] KIM S, YOO C D, NOWOZIN S, et al. Image segmentation usinghigher-order correlation clustering [J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(9): 1761-1774.

[3] JIANG Yizhang, CHUNG F L, WANG Shitong, et al. Collaborative fuzzy clustering from multiple weighted views [J]. IEEE transactions on cybernetics, 2015, 45(4): 688-701.

[4] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms [M]. USA: Springer Science & Business

Media, 2013: 155-201.

[5] KRISHNAPURAM R, KELLER J M. A possibilistic approach to clustering[J]. IEEE transactions on fuzzy systems, 1993, 1(2): 98-110.

[6] KARAYIANNIS N B. MECA: maximum entropy clustering algorithm[C]//Proceedings of the Third IEEE Fuzzy Systems Conference. Orlando, USA: IEEE, 1994: 630-635.

[7] PAN S J, YANG Qiang. A survey on transfer learning[J]. IEEE transactions on knowledge and data engineering, 2010, 22(10): 1345-1359.

[8] DENG Zhaohong, CHOI K S, JIANG Yizhang, et al. Generalized hidden-mapping ridge regression, knowledge-leveraged inductive transfer learning for neural networks, fuzzy systems and kernel methods[J]. IEEE transactions on cybernetics, 2014, 44(12): 2585-2599.

[9] DENG Zhaohong, JIANG Yizhang, CHOI K S, et al. Knowledge-leverage-based TSK fuzzy system modeling[J]. IEEE transactions on neural networks and learning systems, 2013, 24(8): 1200-1212.

[10] ZHI Xiaobin, FAN Jiulun, ZHAO Feng. Fuzzy linear discriminant analysis-guided maximum entropy fuzzy clustering algorithm[J]. Pattern recognition, 2013, 46(6): 1604-1615.

[11] DAI Wenyuan, YANG Qiang, XUE Guirong, et al. Self-taught clustering[C]//Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008: 200-207.

[12] JIANG Wenhao, CHUNG F L. Transfer spectral clustering [M]//FLACH P A, DE BIE T, CRISTIANINI N. Machine Learning and Knowledge Discovery in Databases. Berlin Heidelberg: Springer, 2012: 789-803.

[13] 钱鹏江, 孙寿伟, 蒋亦樟, 等. 知识迁移极大熵聚类算法[J]. 控制与决策, 2015, 30(6): 1000-1006.
QIAN Pengjiang, SUN Shouwei, JIANG Yizhang, et al. Knowledge transfer based maximum entropy clustering[J]. Control and decision, 2015, 30(6): 1000-1006.

[14] PEDRYCZ W, RAI P. Collaborative clustering with the use of Fuzzy C-Means and its quantification[J]. Fuzzy sets and systems, 2008, 159(18): 2399-2427.

[15] GU Quanquan, ZHOU Jie. Learning the shared subspace for multi-task clustering and transductive transfer classification[C]//Proceedings of the Ninth IEEE International Conference on Data Mining. Miami, USA: IEEE, 2009: 159-168.

[16] GU Quanquan, ZHOU Jie. Co-clustering on manifolds [C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2009: 359-368.

[17] RANDEN T. Brodatz texture[EB/OL]. [2015-12-14]. http://www.ux.uis.no/~tranden/brodatz.html.

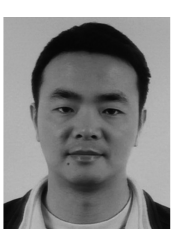
[18] DENG Zhaohong, CHOI K S, CHUNG F L, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information[J]. Pattern recognition, 2010, 43(3): 767-781.

[19] KYRKI V, KAMARAINEN J K, KÄLVIÄINEN H. Simple Gabor feature space for invariant object recognition[J]. Pattern recognition letters, 2004, 25(3): 311-318.

作者简介:

程旸,男,1991 年生,硕士研究生,主要研究方向为人工智能、模式识别、数据挖掘。

蒋亦樟,男,1988 年生,博士,讲师,主要研究方向为人工智能、模式识别、模糊系统。



钱鹏江,男,1979 年生,副教授,博士,主要研究方向为模式识别、医学图像处理、大数据。