DOI:10.11992/tis.201602003
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170227.1758.006.html

基于极大熵的知识迁移模糊聚类算法

陈爱国[1,2],王士同[1]
(1.江南大学 数字媒体学院,江苏 无锡 214122; 2.香港理工大学 计算机系,香港 九龙 999077)

摘　要:针对传统的聚类算法在样本数据量不足或样本受到污染情况下的聚类性能下降问题,在经典的极大熵聚类算法(MEKTFCA)的基础上,提出了一种新的融合历史聚类中心点和历史隶属度这两种知识的基于极大熵的知识迁移模糊聚类算法。该算法通过学习由源域总结出来的有益历史聚类中心和历史隶属度知识来指导数据量不足或受污染的目标域数据的聚类任务,从而提高了聚类性能。通过一组模拟数据集和两组真实数据集构造的迁移场景上的实验,证明了该算法的有效性。

关键词:知识迁移;极大熵;聚类算法;极大熵聚类;模糊聚类
中图分类号:TP274　**文献标志码**:A　**文章编号**:1673-4785(2017)12-0095-09

中文引用格式:陈爱国,王士同.基于极大熵的知识迁移模糊聚类算法[J]. 智能系统学报, 2017, 12(1): 95-103.
英文引用格式:CHEN Aiguo, WANG Shitong. A maximum entropy-based knowledge transfer fuzzy clustering algorithm[J]. CAAI transactions on intelligent systems, 2017, 12(1): 95-103.

A maximum entropy-based knowledge transfer fuzzy clustering algorithm

CHEN Aiguo[1,2], WANG Shitong[1]
(1. School of Digital Media, Jiangnan University, Wuxi 214122, China; 2. Department of Computing, Hong Kong Polytechnic University, Kowloon 999077, China)

Abstract: To address the issue of clustering performance degradation when traditional clustering algorithms are applied to insufficient and/or noisy data, a maximum entropy-based knowledge transfer fuzzy clustering algorithm is proposed. This improves the classical maximum entropy clustering algorithm for target domains by leveraging two kinds of knowledge from the source domain, i.e., historical clustering centers and historical degree of membership, into the objective function proposed for clustering insufficient and/or noisy target data. The effectiveness of the proposed algorithm is demonstrated by experiments on several synthetic and two real datasets.

Keywords: knowledge transfer; maximum entropy; clustering algorithms; maximum entropy clustering; fuzzy clustering

聚类是一种常用的数据分析方法,在人工智能、模式识别和机器学习等领域[1-3]一直受到广泛关注。聚类作为一种无监督的数据分析技术,通过数据之间的疏密程度,将数据划分到不同的数据簇中,使得簇内的数据之间关系比较紧密,而簇之间的数据关系比较疏远。根据聚类使用方法的不同,将聚类分成常见的一些类别:基于划分的聚类算法[4-7]、基于层次的聚类算法[8-9]、基于密度的聚类算法[10-11]、基于图论的聚类算法[12]等。这些聚类算法在针对特定的数据集进行聚类时,通常能获得理想的聚类效果。但这些聚类性能的有效获得都离不开一个必要前提,那就是进行聚类时的数据必须是充分的,换句话说,这些聚类算法不适合处理数据不充分的情况。

但在实际生产、生活中,数据不充分的情况或数据受到污染的情况往往普遍存在。例如,在一个新领域收集数据之初,数据往往是不充足的。又或者由于受到硬件设备的不稳定性或环境等一些因

收稿日期:2016-02-04.　网络出版日期:2017-02-27.
基金项目:国家自然科学基金项目(61272210); 江苏省自然科学基金项目(BK20130155).
通信作者:陈爱国. E-mail:agchen@jiangnan.edu.cn.

素的影响,可能会采集到受噪声干扰的失真数据。对于不充足的数据和受到污染的数据进行聚类分析时,如果直接采用传统聚类方法,往往会造成聚类结果的不理想,甚至有时会出现聚类失效的结果。

如何有效解决数据量不足和数据受污染情况下的数据聚类性能问题,是近年来研究工作者的方向之一。其中,知识迁移[13]机制的引入是一种有效手段。知识迁移机制是指将历史数据(也称为源域)中提炼的有益知识应用到对当前数据(也称为目标域)聚类任务的指导,用于提高当前数据的聚类结果。历史数据与当前数据之间既存在着联系,也存在着明显的差别。目前,应用知识迁移机制来提高聚类性能的代表性算法有:在多任务中使用共享子空间进行聚类的 LSSMTC 算法[14]、使用 K 均值组合方法的 CombKM 算法[14]、使用自学习迁移机制的 STC 聚类算法[15]、使用特征和样本协同机制的 Co-clustering 聚类算法[16]及迁移谱聚类 TSC 算法[17]。这些基于知识迁移的聚类算法虽然提高了一定的聚类性能,但离实际应用还有一定差距,且这些聚类算法在进行聚类任务时需要完整的历史数据集,这在一些特殊场合,如保密需要,完整的历史数据是不可能获得的。所以,研究一种有隐私保护的高效迁移聚类算法具有必要性和实用性。

本文在经典的 MECA 聚类算法的基础上,通过对 MECA 算法的目标函数进行改造,使其具有学习历史知识的能力,进而提高算法在样本量不充分或受到污染情况下聚类的性能。

1 经典极大熵聚类算法

在传统 C 均值聚类算法的基础上,通过引入具有明确物理含义的熵,产生出了具有简洁的数学表达式和明确物理含义特点的极大熵模糊聚类算法。极大熵模糊聚类算法有很多种不同的表达形式,其中文献[6-7]给出的是经典的极大熵模糊聚类算法,其具体表述如下。

假设样本空间 $X=\{x_i \mid x_i \in \mathbf{R}^d, i=1,2,\cdots,N\}$,其中,$N$ 表示样本点的个数,$\mathbf{R}$ 是实数集,d 表示样本点的维数。该样本包含 $C(1<C<N)$ 个不同的类别,则经典的极大熵聚类算法(MECA)的目标函数可表示为

$$J(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2+\alpha\sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}\ln\mu_{ij},$$

$$\mu_{ij}\in[0,1],\sum_{j=1}^{C}\mu_{ij}=1,1\leqslant i\leqslant N \qquad (1)$$

式中:x_i 表示第 i 个样本点;v_j 表示第 j 类中心点;μ_{ij} 表示第 i 个样本点隶属于第 j 类的程度;$\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2$ 表示第 i 个样本点与第 j 类中心点的距离;α 是正则化系数,由 μ_{ij} 构成隶属度矩阵 $\boldsymbol{U}\in\mathbf{R}^{N\times C}$,由 $\boldsymbol{v}_j$ 构成中心点矩阵 $\boldsymbol{V}\in\mathbf{R}^{d\times C}$。

采用拉格朗日条件极值优化方法解得式(1)的最优聚类中心 $\boldsymbol{V}$ 和隶属度 $\boldsymbol{U}$ 的迭代公式为

$$\boldsymbol{v}_j=\frac{\sum_{i=1}^{N}\mu_{ij}\boldsymbol{x}_i}{\sum_{i=1}^{N}\mu_{ij}},\qquad j=1,2,\cdots,C \qquad (2)$$

$$\mu_{ij}=\frac{\exp\left(-\frac{\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2}{\alpha}\right)}{\sum_{k=1}^{C}\exp\left(-\frac{\parallel \boldsymbol{x}_i-\boldsymbol{v}_k\parallel^2}{\alpha}\right)},\qquad i=1,2,\cdots,N$$

$$j=1,2,\cdots,C \qquad (3)$$

根据上述推导,总结出 MECA 算法的具体步骤如下:

输入 数据集 $\boldsymbol{X}$,分类数 C,正则化系数 α,最大迭代次数 T,终止阈值 ε。

输出 最优隶属度 $\boldsymbol{U}$ 和聚类中心 $\boldsymbol{V}$。

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

2)根据式(2)和 1)的隶属度矩阵 $\boldsymbol{U}(t)$ 获得新的类中心 $\boldsymbol{V}(t)$。

3)根据式(3)和 2)获得的类中心 $\boldsymbol{V}(t)$ 计算得新的隶属度 $\boldsymbol{U}(t+1)$。

4)当 $\parallel \boldsymbol{U}(t+1)-\boldsymbol{U}(t)\parallel<\varepsilon$ 或者 $t>T$ 时算法终止,否则跳转到 2)。

通过观察 MECA 算法的具体步骤可以看出,原始的 MECA 算法不具有知识迁移的能力。

MECA 算法在数据量充足时,可以使用上述迭代过程获得有效的类中心和隶属度。但当样本量不充足或样本被污染的情况下,直接使用 MECA 算法获得的聚类中心往往严重偏离实际聚类中心,甚至有时会出现聚类失效的情况。因此,在数据量不足或数据受到污染情况下,研究有效的聚类算法,具有必要性和实际价值。

2 基于极大熵的知识迁移模糊聚类

在知识迁移理论[13]中,当源域数据和目标域数据既存在一定的相关性,同时又存在着明显的差异时,可通过对源域有益知识的充分利用来指导目标域任务更好地完成。本文尝试通过将数据量充分的源域知识迁移至数据量不足或被污染的目标域的聚类任务中,来提高目标域的聚类性能。

为了实现源域知识到目标域迁移的目的,需要解决 3 个核心问题:

1)迁移什么知识;

2)如何迁移;

3)什么时候迁移。

首先,解决源域到目标域迁移什么知识的问题。在基于划分的聚类算法中,隶属度和聚类中心点是对聚类结果具有决定性作用的两个因素。故本文选择隶属度和聚类中心点作为被迁移的知识。

其次,需要解决如何才能实现将源域的隶属度和聚类中心点知识迁移到目标域的聚类任务中的问题。我们通过以下两个规则来实现。

1)隶属度重要程度受约束规则

该规则对应的公式为

$$\min J_{\mathrm{KT}_1}(\boldsymbol{X},\boldsymbol{U},\boldsymbol{V},\hat{\boldsymbol{U}}) = \lambda\sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2 + (1-\lambda)\sum_{i=1}^{N}\sum_{j=1}^{C}\hat{\mu}_{ij}\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2,\quad \lambda\in[0,1] \tag{4}$$

式中:$\boldsymbol{x}_i$ 表示目标域中第 i 个样本点;$\boldsymbol{v}_j$ 表示目标域中第 j 类中心点;μ_{ij} 表示目标域中第 i 个样本点隶属于第 j 类的程度;$\hat{\mu}_{ij}$ 为目标域中第 i 个样本点相对于源域中第 j 类中心点的历史隶属度;λ 为隶属度重要程度受约束规则的平衡因子。通过隶属度重要程度受约束规则来调整性地学习源域迁移过来的历史隶属度知识。平衡因子 λ 控制着源域的历史隶属度和目标域的隶属度对最终聚类结果的影响程度。当 $\lambda\to1$ 时,说明迁移的历史隶属度的可靠程度差,目标域的聚类结果更多地受到目标域的隶属度的影响。当 $\lambda\to0$ 时,说明迁移的历史隶属度具有很高的可借鉴性,目标域的聚类结果更多地受到迁移历史隶属度的影响。

2)聚类中心点变化最小规则

该规则的公式为

$$\min J_{\mathrm{KT}_2}(\boldsymbol{V},\hat{\boldsymbol{V}}) = \beta\sum_{j=1}^{C}\parallel \boldsymbol{v}_j-\hat{\boldsymbol{v}}_j\parallel^2,\qquad \beta\geqslant0 \tag{5}$$

式中:$\boldsymbol{v}_j$ 代表目标域的第 j 类中心点,$\hat{\boldsymbol{v}}_j$ 代表源域的第 j 类的历史类中心点;β 表示类中心点变化最小规则的平衡因子;C 是聚类数目。该规则实现的是源域聚类中心点知识的迁移。通过该规则的使用确保:当目标函数产生最优化解时,目标域的聚类中心点在一定程度上与源域的中心点保持一致,并通过平衡因子 β 来控制保持一致的程度。当 $\beta\to0$ 时,目标域的聚类中心点与源域的聚类中心点需保持一致性程度小,此时说明源域的聚类中心点知识不可靠。当 $\beta\to\infty$ 时,目标域的聚类中心点与源域的聚类中心点需保持一致性程度大,此时说明源域的聚类中心点知识可靠性高。

根据上述分析,针对经典 MECA 算法不具有知识迁移能力的不足,本文在 MECA 算法的基础上结合上述两个规则,提出基于极大熵知识迁移模糊聚类算法,即 MEKTFCA 算法。该算法的原理如图 1。

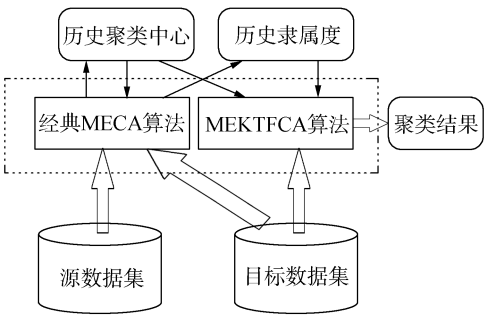


图 1　MEKTFCA 算法原理图

Fig.1　Overall framework of MEKTFCA algorithm

MEKTFCA 算法首先对源数据集,通过经典的 MECA 算法获得历史聚类中心,然后根据目标数据集和所获得的历史聚类中心,通过再次使用经典的 MECA 算法获得历史隶属度,最后通过 MEKTFCA 算法和历史聚类中心及历史隶属度获得最终的聚类结果。

融入了隶属度重要程度受约束规则和聚类中心点变化最小规则得到的 MEKTFCA 算法的具体目标函数为

$$J_{\mathrm{MEKTFCA}}(\boldsymbol{X},\boldsymbol{V},\hat{\boldsymbol{V}},\boldsymbol{U},\hat{\boldsymbol{U}}) = J_{\mathrm{KT}_1}(\boldsymbol{X},\boldsymbol{U},\boldsymbol{V},\hat{\boldsymbol{U}}) + J_{\mathrm{KT}_2}(\boldsymbol{V},\hat{\boldsymbol{V}}) + \alpha\sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}\ln\mu_{ij},\quad \mu_{ij}\in[0,1];\ \sum_{j=1}^{C}\mu_{ij}=1,\ \forall i=1,2,\cdots,N \tag{6}$$

其中

$$J_{\mathrm{KT}_1}(\boldsymbol{X},\boldsymbol{U},\boldsymbol{V},\hat{\boldsymbol{U}}) = \lambda\sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2 + (1-\lambda)\sum_{i=1}^{N}\sum_{j=1}^{C}\hat{\mu}_{ij}\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2 \tag{7}$$

$$J_{\mathrm{KT}_2}(\boldsymbol{V},\hat{\boldsymbol{V}}) = \beta\sum_{j=1}^{C}\parallel \boldsymbol{v}_j-\hat{\boldsymbol{v}}_j\parallel^2 \tag{8}$$

观察式(6)可以看出,MEKTFCA 算法从本质上来说是由 3 部分组成。第 1 部分是融入了对历史隶属度知识的学习的 $J_{\mathrm{KT}_1}(\boldsymbol{X},\boldsymbol{U},\boldsymbol{V},\hat{\boldsymbol{U}})$ 项。通过该项的引入可以使用历史隶属度知识对目标域的聚类任务进行指导。第 2 部分是融入了对历史类中心点知识学习的 $J_{\mathrm{KT}_2}(\boldsymbol{V},\hat{\boldsymbol{V}})$ 项。通过该项的引入可以更有效地帮助目标域聚类任务的执行。第 3 部分是原

MECA 算法的正则化熵项。同时,根据式(6)~(8)可以发现,当隶属度重要程度受约束规则的平衡因子 $\lambda=1$ 而且聚类中心点变化最小规则的平衡因子 $\beta=0$ 这种特殊情况时,MEKTFCA算法就退化为经典的 MECA 算法。MEKTFCA 算法的本质是,通过调节平衡因子 λ 和 β 的大小,来调整历史隶属度和历史类中心点对当前聚类任务的影响,从而改善由于数据量不足和数据被污染情况下直接采用 MECA 算法造成聚类结果不理想的情况。

2.1 参数求解

式(6)的参数求解问题,即为在有约束条件下求解最优的参数使得目标函数值最小。与 MECA 求最优解方法相同,我们采用拉格朗日乘子法进行求解,首先构造拉格朗日函数表达式,即

$$L=\lambda\sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2+(1-\lambda)\sum_{i=1}^{N}\sum_{j=1}^{C}\hat{\mu}_{ij}\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2+\alpha\sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}\ln\mu_{ij}+\beta\sum_{j=1}^{C}\parallel \boldsymbol{v}_j-\hat{\boldsymbol{v}}_j\parallel^2+\sum_{i=1}^{N}\eta_i\left(\sum_{j=1}^{C}\mu_{ij}-1\right)\tag{9}$$

式中 η_i 为 Lagrange 乘子。

根据 $\frac{\partial L}{\partial\mu_{ij}}=0$ 得

$$\mu_{ij}=\exp\left(-1-\frac{\eta_i+\lambda\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2}{\alpha}\right)\tag{10}$$

因为有约束条件

$$\sum_{j=1}^{C}\mu_{ij}=1\tag{11}$$

将式(10)代入式(11)可得

$$\exp\left(-\frac{\eta_i}{\alpha}\right)=\frac{\exp(1)}{\sum_{k=1}^{C}\exp\left(-\frac{\lambda\parallel \boldsymbol{x}_i-\boldsymbol{v}_k\parallel^2}{\alpha}\right)}\tag{12}$$

将式(12)代入式(10)可得隶属度的迭代公式为

$$\mu_{ij}=\frac{\exp\left(-\frac{\lambda\parallel \boldsymbol{x}_i-\boldsymbol{v}_j\parallel^2}{\alpha}\right)}{\sum_{k=1}^{C}\exp\left(-\frac{\lambda\parallel \boldsymbol{x}_i-\boldsymbol{v}_k\parallel^2}{\alpha}\right)}\tag{13}$$

根据 $\frac{\partial L}{\partial\boldsymbol{v}_j}=0$ 解得聚类中心点迭代公式为

$$\boldsymbol{v}_j=\frac{\sum_{i=1}^{N}[\lambda\mu_{ij}+(1-\lambda)\hat{\mu}_{ij}]\boldsymbol{x}_i+\beta\hat{\boldsymbol{v}}_j}{\sum_{i=1}^{N}[\lambda\mu_{ij}+(1-\lambda)\hat{\mu}_{ij}]+\beta}\tag{14}$$

2.2 算法步骤

根据上述的推导过程所获得的隶属度和聚类中心点的迭代公式,给出 MEKTFCA 算法的具体步骤如下。

输入 历史类中心点 $\hat{\boldsymbol{v}}_j$,目标数据集 X,分类数 C,平衡因子 λ、β,正则化系数 α,最大迭代次数 T,终止阈值 ε。

输出 最优隶属度 $\boldsymbol{U}$ 和聚类中心 $\boldsymbol{V}$。

1)根据式(3)计算历史隶属度 $\hat{\mu}_{ij}$。

2)初始化迭代计数器 $t=0$。

3)根据式(14)计算得到新的聚类中心 $\boldsymbol{V}(t)$。

4)根据式(13)计算得到新的隶属度矩阵 $\boldsymbol{U}(t+1)$。

5)当 $\parallel \boldsymbol{U}(t+1)-\boldsymbol{U}(t)\parallel<\varepsilon$ 或者 $t>T$ 时算法终止,否则跳转到 3)。

以上算法步骤同时回答了实现知识迁移中的第 3 个核心问题:什么时候迁移。通过算法步骤可以看到,在算法不断迭代过程中,隶属度和中心点的迭代公式中使用到了历史隶属度知识和历史类中心点知识。从而在算法的迭代过程中实现了知识的迁移。

3 实验结果及分析

3.1 实验设置

为验证本文所提 MEKTFCA 算法的有效性,将构造一组模拟数据集和两组真实数据集作为实验所使用的迁移场景。同时,选择 6 种相关算法作为对比算法,对它们的聚类性能进行比较。这 6 种算法为:在多任务中使用共享子空间进行聚类的 LSSMTC 算法[14],使用 K 均值组合方法的 CombKM 算法[14],使用自学习迁移机制的 STC 聚类算法[15],使用特征和样本协同机制的 Co-clustering 聚类算法[16],迁移谱聚类 TSC 算法[17]以及经典的 MECA 算法[6-7]。

为了对聚类算法的结果进行客观比较,本文采用统一的 NMI[18](normalized mutual information)和 RI[19](rand index)两种指标对实验结果进行评价。NMI 的计算公式为

$$\mathrm{NMI}=\frac{\sum_{p=1}^{c}\sum_{q=1}^{c}N_{p,q}\log\left(\frac{N\cdot N_{p,q}}{N_p\cdot N_q}\right)}{\sqrt{\left[\sum_{p=1}^{c}N_p\log\left(\frac{N_p}{N}\right)\right]\cdot\left[\sum_{q=1}^{c}N_q\log\left(\frac{N_q}{N}\right)\right]}}$$

式中:N 代表数据集的样本数目;N_p 代表聚类到 p 类的样本数目;N_q 代表类标签为 q 的样本数目;$N_{p,q}$ 代表同时聚类到 p 类和类标签为 q 的样本数目。RI 评价指标的计算公式为

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

式中:N_{00}代表拥有不同类标签的两个样本聚类到不同类别中的个数;N_{11}代表拥有相同类标签的两个样本聚类到相同类别中的个数。上述两种评价指标值的变化范围都为[0,1],且值越大,代表其算法的聚类性能越好。

在 MEKTFCA 算法中,涉及两个平衡因子 λ 和 β 如何取值的问题。本文采用网格搜索进行遍历寻优,其寻优的范围及其他对比算法的参数设置值如表 1 所示。

表 1　算法参数设置值

Table 1　Parameter sets for algorithms

算法	参数设置值
LSSMTC	共享子空间的维度 $l=1$,任务数 $m=2$,正则化参数 $\lambda=0.5$
CombKM	K 等于聚类数
STC	平衡参数 $\lambda=3$
Co-clustering	特征聚类数 $m=2$
TSC	平衡因子 $\lambda=1$,步长 step=1
MECA	熵正则化参数 $\gamma\in\{0.1:0.2:1\}\cup\{2:1:10\}\cup\{20:10:100\}$
MEKTFCA	熵正则化参数 $\alpha\in\{0.1:0.2:1\}\cup\{2:1:10\}\cup\{20:10:100\}$ 平衡因子 $\lambda\in\{0:0.1:1\}$ 平衡因子 $\beta\in\{0:0.2:1\}\cup\{2:1:10\}\cup\{20:10:100\}$

本实验所采用的环境是:Intel i7-5600U 2.60 GHz 8 GB RAM;Windows 8 64 bit;MATLAB R2012b。实验所列结果数据均是在运行 10 次后求得的平均值。

3.2　模拟数据集实验结果和分析

该实验是通过构造的模拟数据集来验证本文所提的 MEKTFCA 算法在样本量不足和受污染情况下聚类算法的有效性。本实验构造了一组源数据集 S 和 3 组目标数据集 T_1、T_2、T_3,其中源数据集和所有的目标数据集均包含 3 类 2 维的数据,这些数据的生成均采用高斯概率分布模型函数,生成时所使用的均值、方差及每个类别包含的样本数如表 2 所示。

构造的源数据集 S 共含有1 500个样本,这个数据集数据量充足,并且能够从该数据集中提取出对目标数据集的聚类具有指导作用的有用知识。

构造的目标数据集 T_1 共含有 90 个样本,只占源数据集 S 总数据量的 6%,用于代表数据量不充足的场景,虽然数据集 T_1 与数据集 S 的均值存在差异,但它们的方差相同,这用于体现迁移学习中的源数据集与目标数据集之间既存在着相似性,同时也存在着一定的差别的情况。

表 2　模拟数据集生成的参数设置

Table 2　Parameter sets to generate synthetic datasets

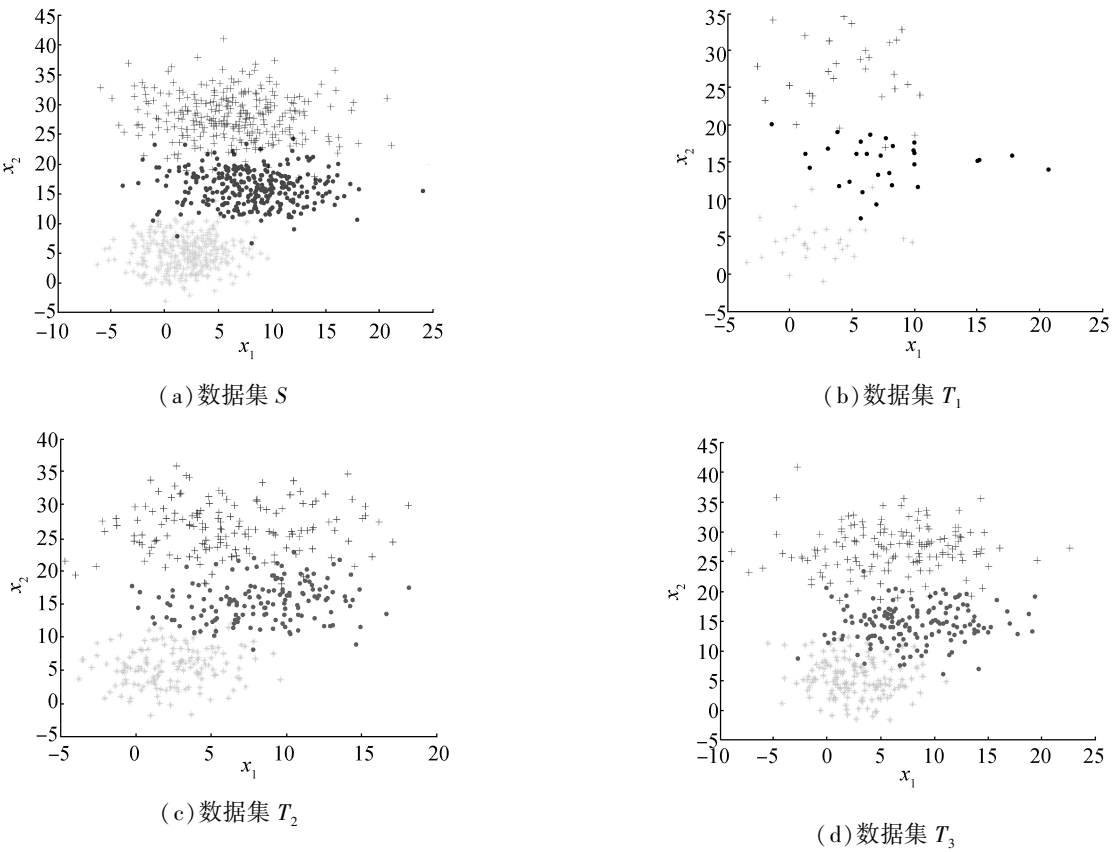
数据集	类别	均值	方差	样本数
源数据集 S	第 1 类	$\begin{bmatrix}2\\5\end{bmatrix}$	$\begin{bmatrix}8&0\\0&8\end{bmatrix}$	500
	第 2 类	$\begin{bmatrix}8\\16\end{bmatrix}$	$\begin{bmatrix}16&0\\0&9\end{bmatrix}$	500
	第 3 类	$\begin{bmatrix}6\\28\end{bmatrix}$	$\begin{bmatrix}25&0\\0&15\end{bmatrix}$	500
目标数据集 T_1	第 1 类	$\begin{bmatrix}2.5\\5.0\end{bmatrix}$	$\begin{bmatrix}8&0\\0&8\end{bmatrix}$	30
	第 2 类	$\begin{bmatrix}8\\15\end{bmatrix}$	$\begin{bmatrix}16&0\\0&9\end{bmatrix}$	30
	第 3 类	$\begin{bmatrix}6\\27\end{bmatrix}$	$\begin{bmatrix}25&0\\0&15\end{bmatrix}$	30
目标数据集 T_2、T_3	第 1 类	$\begin{bmatrix}2.5\\5.0\end{bmatrix}$	$\begin{bmatrix}8&0\\0&8\end{bmatrix}$	150
	第 2 类	$\begin{bmatrix}8\\15\end{bmatrix}$	$\begin{bmatrix}16&0\\0&9\end{bmatrix}$	150
	第 3 类	$\begin{bmatrix}6\\27\end{bmatrix}$	$\begin{bmatrix}25&0\\0&15\end{bmatrix}$	150

构造的目标数据集 T_2 共含有 450 个样本,占源数据集 S 总数据量的 30%。用目标数据集 T_2 来代表目标数据量充分的场景。

构造的目标数据集 T_3 与目标数据集 T_2 的均值、方差和数据量完全相同。不同的是,数据集 T_3 在数据集 T_2 的基础上增加了方差为 2、均值为 0 的高斯噪声,用于代表数据量充分但受到了噪声污染的场景。

上述构造的 4 组模拟数据集的数据分布如图 2 所示。

在上述构造出来的各种迁移场景下,运行 MEKTFCA 算法和 6 种对比算法,得到的实验结果如表 3 所示。因为 TSC 算法要求样本的维数要大于聚类数目,而在模拟数据集上不满足此条件,所以在表 3 中我们使用“—”来表示此算法无法运行。



(a)数据集 S (b)数据集 T_1

(c)数据集 T_2 (d)数据集 T_3

图 2 模拟数据集分布图

Fig.2 Data distribution of synthetic datasets

表 3 模拟数据集上 7 种聚类算法的性能对比

Table 3 Clustering performance of seven algorithms on synthetic datasets

场景	评价指标	LSSMTC	CombKM	STC	Co-lustering	TSC	MECA	MEKTFCA
S-T_1	NMI-mean	0.674 1	0.759 4	0.155 6	0.737 2	—	0.737 2	0.802 4
	NMI-std	0.036 3	$1.170\ 3\times10^{-16}$	0	$1.170\ 3\times10^{-16}$	—	$1.170\ 3\times10^{-16}$	0.002 2
	RI-mean	0.867 4	0.902 4	0.605 2	0.901 4	—	0.901 4	0.917 5
	RI-std	0.0194	$1.170\ 3\times10^{-16}$	$1.170\ 3\times10^{-16}$	$2.340\ 6\times10^{-16}$	—	$2.340\ 6\times10^{-16}$	0.007 1
S-T_2	NMI-mean	0.734 3	0.787 6	0.169 4	0.771 9	—	0.764 9	0.834 8
	NMI-std	0.032 4	0	$2.925\ 7\times10^{-17}$	0.009 9	—	$7.401\ 5\times10^{-17}$	0.001 9
	RI-mean	0.905 9	0.921 5	0.687 9	0.919 5	—	0.916 8	0.940 1
	RI-std	0.012 8	0	$1.170\ 3\times10^{-16}$	0.005 3	—	0	0.001 1
S-T_3	NMI-mean	0.734 6	0.773 7	0.126 1	0.761 5	—	0.774 6	0.816 2
	NMI-std	0.028 5	$1.170\ 3\times10^{-16}$	0	0.009 4	—	$1.170\ 3\times10^{-16}$	$1.170\ 3\times10^{-16}$
	RI-mean	0.899 5	0.914 7	0.587 8	0.911 1	—	0.914 8	0.937 2
	RI-std	0.010 8	$1.170\ 3\times10^{-16}$	0	0.002 6	—	0	$1.170\ 3\times10^{-16}$

观察表 3 的实验结果可以得出如下结论：

1）源数据集是 S，目标数据集是 T_1，它们所代表的是数据量不足的场景。从该场景中的实验结果可以看出，没有使用任何迁移机制的 MECA 算法，在数据量不充分的情况下，其聚类性能要明显比使用了知识迁移的 MEKTFCA 算法差。因为 MECA 算法对数据量不充足的数据集进行聚类时，产生的聚类中心将会严重偏离实际的聚类中心，进而导致最终的聚类结果不理想。而 MEKTFCA 算法由于引入了知识迁移机制，在聚类过程中，通过学

习由源数据集提取的历史类中心和历史隶属度中的有用知识,有效地改善了由于数据量不足而导致的聚类中心偏移的问题,最终提高了聚类性能。对于另外的 4 种对比算法,从它们的实验结果可以看出,虽然这些算法都使用了不同的机制来提高聚类结果,但由于算法本身的限制,在样本量不充足的场景下,其聚类结果不理想。

2) 源数据集是 S,目标数据集是 T_2,它们所构成的是数据量较充足且无污染的场景。从该场景中的实验结果可以看出,由于目标数据集 T_2 的数据量较充分,因此所有 6 种算法均取得比较理想的聚类结果。这也进一步说明了传统的聚类算法进行有效聚类的前提条件是数据量要充分。在这 6 种聚类算法中,由于 MEKTFCA 算法对源数据集所形成的历史中心点和历史隶属度双重有益知识进行了学习,它的聚类结果要优于其他 5 种算法。

3)源数据集是 S,目标数据集是 T_3,它们构成的是数据量较充分但受到污染的场景。从该场景中的实验结果可以看出,当数据受到污染时,其余 5 种对比算法的聚类性能要明显差于本文提出的 MEKTFCA算法。因为数据集 T_3 受到了污染,从而导致数据产生失真,数据集的聚类中心产生了显著的偏移,数据类别之间的界限变得模糊,最终使得这五种聚类算法的聚类性能不够理想。MEKTFCA 算法在进行聚类时,通过调整两个平衡因子的权重来学习源数据集中有益的历史中心点知识和历史隶属度知识,这种迁移知识的学习机制提高了 MEKTFCA 算法在数据被污染情况下的聚类效果。这也使得 MEKTFCA 算法具有一定的抗噪性能。

3.3　文本数据集实验结果和分析

第 2 个实验基于真实的文本数据集 20 Newsgroups(20 NG)[20]构造的目标数据样本量不足的迁移场景来对 MEKTFCA 算法的有效性进行进一步验证。20 NG 数据集包含大约 20 000 条新闻组信息,均匀地分布在 20 个不同的集合中。我们选择这 20 个集合中的 4 个来生成两组实验用的迁移场景:“comp VS sci”和“rec VS talk”。这两组迁移场景的数据集构成如表 4 所示。

表 4　文本迁移场景数据集构成

Table 4　Text transfer scenes structures

迁移场景	数据集	数据个数	维数
comp VS sci	历史	1 500	350
	目标	150	350
rec VS talk	历史	1 500	350
	目标	150	350

这两组迁移场景所使用的数据集的来源见表 5 所示。在构造这两组迁移场景时,源数据集和目标数据集之间有一定的相似性,同时又有明显的不同。如构造的“comp VS sci”迁移场景第 1 类的源数据集和目标数据集都来自同样的“comp”大类,但它们的子类是完全不同的。第 2 类的源数据集和目标数据集与此类似。因为 20 NG 数据集的原始数据的维数较大,因此实验之前我们使用工具 BOW[21] 对其进行了降维处理。本文所提的 MEKTFCA 算法及其 6 种对比算法在此数据集上的实验结果如表 6。

表 5　文本迁移场景的数据集来源

Table 5　Text transfer scenes data sources

迁移场景	类别	源数据集	目标数据集
comp VS sci	1	comp.os.ms-windows.misc	comp.windows.x
	2	sci.crypt	sci.space
rec VS talk	1	rec.autos	rec.sport.baseball
	2	talk.politics.guns	talk.politics.mideast

表 6　文本迁移场景中 7 种聚类算法性能对比

Table 6　Clustering performance of seven algorithms on text transfer scenes

迁移场景	评价指标	LSSMTC	CombKM	STC	Co-lustering	TSC	MECA	MEKTFCA
comp VS sci	NMI-mean	0.390 2	0.108 1	0.275 2	0.281 0	0.860 2	0.689 1	1
	NMI-std	0.246 3	0.273 2	0.161 7	0.295 8	$1.170\ 3\times10^{-16}$	0	0
	RI-mean	0.648 7	0.547 8	0.645 7	0.610 0	0.964 6	0.863 2	1
	RI-std	0.148 9	0.145 8	0.100 2	0.161 0	$2.340\ 6\times10^{-16}$	$1.170\ 3\times10^{-16}$	0
VS talk	NMI-mean	0.027 9	0.165 1	0.188 8	0.025 1	0.822 6	0.066 6	0.905 3
	NMI-std	$7.314\ 2\times10^{-18}$	0.108 4	0.007 5	0.008 8	$1.170\ 3\times10^{-16}$	0.077 4	0.026 0
	RI-mean	0.496 7	0.533 9	0.600 7	0.496 7	0.947 7	0.505 6	0.971 0
	RI-std	0	0.042 1	0.002 0	$2.829\ 8\times10^{-5}$	0	0.027 1	0.010 2

观察表 6 的实验结果可以得出如下结论：

1）MEKTFCA 算法和 TSC 算法在迁移场景“comp VS sci”上的聚类结果较其他 5 种算法的聚类结果优势明显。其他 5 种算法中，MECA 算法的聚类结果最好，LSSMTC 算法次之，Co-clustering 算法、STC 算法和 CombKM 算法的聚类结果明显差于其他算法。究其原因，主要是由于在此构造的真实场景上 MEKTFCA 算法和 TSC 算法的迁移学习机制比其他算法更有效。

2）MEKTFCA 算法和 TSC 算法在构造的迁移场景“rec VS talk”上的聚类结果具有同样明显的优势，而其他 5 种算法的聚类结果却差很多。MECA 算法因为在此迁移场景下，目标数据集的样本量不足，且没有引入任何迁移学习机制，导致最终的聚类结果较差。STC 算法、CombKM 算法、LSSMTC 算法和 Co-clustering 算法尽管都使用了不同的迁移学习机制，但在此迁移场景下的效果不明显，所以它们的聚类结果也明显较差。而 MEKTFCA 算法和 TSC 算法在此迁移场景上的知识迁移效果明显，故取得较好的聚类结果。

3）进一步观察表 6 的实验结果可以发现，MEKTFCA 算法在所构造的两组迁移场景上的聚类性能都明显高于经典的 MECA 算法，这主要是因为 MEKTFCA 算法是在 MECA 算法的基础上通过改变两个平衡因子的大小来调整对历史中心点和历史隶属度知识的学习权重。因为对迁移知识的有效学习，所以保证了 MEKTFCA 算法的聚类效果始终要比 MECA 算法的聚类效果好。

3.4 入侵检测数据集实验结果和分析

最后一个实验使用的是真实的入侵检测数据集 KDDCup99[22]，使用该数据集形成一个目标数据集样本量不足的场景，通过将 MEKTFCA 算法应用到该场景，再次验证该算法的有效性。KDDCup99 数据集来源于美国林肯实验室建立的一个模拟网络环境中收集的网络连接和审计数据。数据集中的数据包含不同类型的用户、各种不同类型的网络流量以及各种类型的网络攻击。该数据集共收集了 9 周时间的数据，其中 7 周时间的数据作为训练数据，约含有5 000 000个网络连接数据，另外 2 周时间的数据作为测试数据，约含有2 000 000个网络连接数据。整个数据集中的训练数据集和测试数据集之间有着不同的概率分布，即整个数据集具有一定的相似性，同时也存在着一定的差异性，满足知识迁移应用的条件。本文基于 KDDCup99 数据集构造的知识迁移场景只选择了数据集中常见的 5 种类型（Normal、smurf、Neptune、satan、ipsweep）作为源数据集和目标数据集的类别，选择每条网络连接记录中的 32 个连续型的特征属性作为样本数据的维度，源数据集的样本来自原始的训练数据集，目标数据集的样本来自原始的测试数据集。目标数据集的样本数量只占源数据集的样本数量的 10%，且样本数量不大，表征的是目标数据集样本量不足的情况。具体的入侵检测迁移场景的数据集的样本个数、数据的维数和类别如表 7 所示。

表 7 入侵检测迁移场景的数据集构成

Table 7 KDDCup99 transfer scenes structures

数据集	样本个数	维数	类别
源数据集	2 000	32	5
目标数据集	200	32	

MEKTFCA 算法和其他 6 种对比算法在该入侵检测数据集知识迁移场景上的执行结果如表 8 所示。

表 8 入侵检测迁移场景上 7 种聚类算法性能对比

Table 8 Clustering performance of seven algorithms on KDDCup99 transfer scene

评价指标	LSSMTC	CombKM	STC	Co-clustering	TSC	MECA	MEKTFCA
NMI-mean	0.402 1	0.383 4	0.397 0	0.302 5	0.284 5	0.409 0	0.756 9
NMI-std	0.045 3	0.008 3	0.061 2	0.126 9	0.023 1	$6.409\ 9\times10^{-17}$	0.023 6
RI-mean	0.508 4	0.497 6	0.498 3	0.492 8	0.447 5	0.514 3	0.902 2
RI-std	0.027 5	0.026 3	0.087 6	0.065 0	0.013 1	0	0.017 2

观察表 8 实验结果可以看出，MEKTFCA 算法在两大性能指标 NMI 和 RI 的均值要明显高于其他 6 种对比算法。与上述两个实验结果分析的原因相同，MEKTFCA 算法由于从源数据集中学习了有益的历史中心点和历史隶属度知识，并对目标数据集的聚类过程形成有效指导，进而使得 MEKTFCA 算法在目标数据集样本量不足的情况下仍能取得比其他 6 种算法更好的聚类性能。这同时也进一步验证了对历史知识的有效学习对当前聚类任务的有效完成具有促进作用。

4 结束语

由于传统聚类算法在数据样本量不足或数据受污染的情况下，其聚类效果往往不理想甚至失效，本文在经典的 MECA 算法的基础上通过引入知识迁移机制，提出了基于极大熵的知识迁移模糊聚类算法，即 MEKTFCA 算法。MEKTFCA 算法通过引入的知识迁移机制使得在对数据量不足或受污染的目标域数据进行聚类时能够有效学习源域中有益的历史中心点和历史隶属度知识，进而提高了最终的聚类结果。通过一组模拟数据集和两组真实

的数据集上的实验结果,可以得出本文所提MEKTFCA 算法较 6 种相关算法在聚类性能上有明显的优越性。本文的创新点主要是,在 MECA 算法的基础了提出了一个新的知识迁移方法,该方法能有效解决实际生活中数据短缺和有噪声数据场景下的聚类性能问题。此外,本文所提的 MEKTFCA 算法也存在着局限性,如在不同应用场景下两个平衡参数如何进行有效确定的问题。这也是我们对该算法进行进一步研究的方向。

参考文献:

[1] CARIOU C, CHEHDI K. Unsupervised nearest neighbors clustering with application to hyperspectral images [J]. IEEE journal of selected topics in signal processing, 2015, 9(6): 1105-1116.

[2] ALI A, BOYACI A, BAYNAL K. Data mining application in banking sector with clustering and classification methods [C]//Proceedings of 2015 International Conference on Industrial Engineering and Operations Management. Dubai, UAE, 2015: 1-8.

[3] LI Shuai, ZHOU Xiaofeng, SHI Haibo, et al. An efficient clustering method for medical data applications [C]// Proceedings of 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent System. Shenyang, China, 2015: 133-138.

[4] LIKAS A, VLASSIS N, VERBEEK J J. The global k-means clustering algorithm[J]. Pattern recognition, 2003, 36(2): 451-461.

[5] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms [M]. New York: Springer, 1981: 43-93.

[6] KARAYIANNIS N B. MECA: maximum entropy clustering algorithm[C]//Proceedings of the 3rd IEEE International Conference on Fuzzy Systems. Orlando, USA, 1994, 1: 630-635.

[7] LI Ruiping, MUKAIDONO M. A maximum-entropy approach to fuzzy clustering[C]//Proceedings of 1995 the 4th IEEE International Conference on Fuzzy System. Yokohama, Japan, 1995, 4: 2227-2232.

[8] ZHANG Tian, RAMAKRISHNAN R, LIVNY M. BIRCH: an efficient data clustering method for very large databases [C]//Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. New York, NY, USA, 1996: 103-114.

[9] GUHA S, RASTOGI R, SHIM K. CURE: an efficient clustering algorithm for large databases[C]//Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. New York, NY, USA, 1998: 73-84.

[10] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C]//Proceeding of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon, USA, 1996: 226-231.

[11] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: ordering Points to Identify the Clustering Structure[C]//Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. Philadelphia, Pennsylvania, USA, 1999: 49-60.

[12] ARIAS-CASTRO E, CHEN Guangliang, LERMAN G. Spectral Clustering based on local linear approximations [J]. Electronic journal of statistics, 2011, 5: 1537-1587.

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

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

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

[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, 2009: 359-368.

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

[18] JING Liping, NG K M, HUANG J Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data [J]. IEEE transactions on knowledge and data engineering, 2007, 19 (8): 1026-1041.

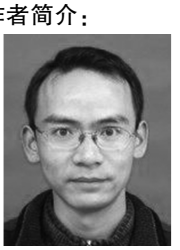
[19] LIU Jun, MOHAMMED J, CARTER J, et al. Distance-based clustering of CGH data[J]. Bioinformatics, 2006, 22(16): 1971-1978.

[20] DAI Wenyuan, XUE Guirong, YANG Qiang, et al. Co-clustering based classification for out-of-domain documents [C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2007: 210-219.

[21] MCCALLUM A K. Bow: a toolkit for statistical language modeling, text retrieval, classification and clustering[EB/OL]. 1996. http://www.cs.cmu.edu/mccallum/bow.

[22] BAY S D, KIBLER D, PAZZANI M J, et al. The UCI KDD archive of large data sets for data mining research and experimentation [J]. ACM SIGKDD explorations newsletter, 2000, 2(2): 81-85.

作者简介:



陈爱国,男,1975 年生,博士研究生,主要研究方向为模式识别与机器学习。

王士同,男,1964 年生,教授,博士生导师,中国离散数学学会常务理事,中国机器学习学会常务理事,主要研究方向为人工智能、模式识别和生物信息。发表学术论文近百篇,其中被 SCI、EI 检索 50 余篇。