具有隐私保护功能的知识迁移聚类算法

陈爱国* 王士同
(江南大学数字媒体学院 无锡 214122)

摘 要：传统聚类算法在数据量不足或数据被污染的场景下聚类效果较差，针对此问题，在经典模糊 C 均值(FCM)技术的基础上，该文提出融合历史类中心和历史隶属度两类知识迁移机制的聚类算法。该算法通过有效利用历史数据中总结得到的辅助知识来指导当前由于数据不足或数据污染带来的聚类困难问题，从而提高聚类效果。同时，由于该算法仅利用历史数据的类中心和隶属度，对历史数据具有隐私保护的优点。通过在模拟数据集和真实数据集上的仿真实验，证明了该算法的有效性。

关键词：知识迁移；隐私保护；聚类算法；模糊 C 均值
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2016)03-0523-09
DOI: 10.11999/JEIT150645

Knowledge Transfer Clustering Algorithm with Privacy Protection

CHEN Aiguo WANG Shitong
(*School of Digital Media*, *Jiangnan University*, *Wuxi* 214122, *China*)

Abstract: For the traditional clustering algorithms efficiency problems in situations of insufficient datasets or datasets with noises, a Knowledge Transfer Clustering Algorithm with Privacy Protection (KTCAPP) is proposed based on the classical Fuzzy C-Means (FCM) technology by leveraging two kinds of knowledge which are the historical class center and the historical class membership. The performance of KTCAPP is enhanced by using auxiliary knowledge from history datasets to guide the current clustering task with insufficient datasets or datasets with noises. In addition, KTCAPP is of good capability of privacy protection because the algorithm only uses the historical class center and the historical class membership which do not expose the raw data. Experiment results show the proposed algorithm is efficient.

Key words: Knowledge transfer; Privacy protection; Clustering algorithm; Fuzzy C-Means (FCM)

1 引言

聚类分析是将一组未知类标签的数据样本按照它们在性质上的亲疏程度划分到由类似性质数据组成的多个类的过程[1−4]。聚类分析作为一种无监督的数据处理技术，是目前机器学习、数据挖掘和人工智能等领域的研究热点之一。目前，聚类分析已被广泛地应用于文本聚类、视频处理、图像分割以及入侵检测等方面。传统的聚类分析算法如基于层次的 BIRCH 算法[5,6]、基于划分的 FCM 算法[7−11]、基于密度的 DBSCAN 算法[12]和基于网格的 WAVE-CLUSTER 算法[13]在处理各自的数据样本时，都能获得不错的聚类性能。但这些聚类分析算法得以有效实施都要基于一个前提条件，即所采用的数据样本数量必须是充分的，数据样本所提供的聚类信息是有效的。

但是实际生产生活中往往存在着这样的情况：一是，在一个新领域的数据样本收集初期，数据样本数据往往不够充分。另一个是，由于传感器的失效或传输线路的干扰等因素的影响使得采集的数据样本受到了污染，这些被污染的数据样本所提供的聚类信息不是全部有效。这些情况的出现，若继续采用传统的聚类分析算法将面临聚类性能不佳甚至失效的可能。

针对上述问题，可以通过在聚类分析算法中引入迁移学习[14−16]机制来解决。迁移学习利用数据样本量充分的历史域来指导数据样本量不充分的目标域。其中历史域和目标域的数据在特征空间或属性方面总体上相近，但也存在一定的差异性。通过从

收稿日期：2015-06-01；改回日期：2015-11-02；网络出版：2015-12-04
*通信作者：陈爱国 agchen@jiangnan.edu.cn
基金项目：国家自然科学基金(61272210)，江苏省杰出青年基金(BK20140001)，江苏省自然科学基金(BK20130155)
Foundation Items: The National Natural Science Foundation of China (61272210), Jiangsu Province Outstanding Youth Fund (BK20140001), Natural Science Foundation of Jiangsu Province (BK20130155)

充足的历史域数据中获得有益的历史知识来指导当前数据量不足或数据被污染的目标域，从而提高当前数据的聚类效果。

近年来，在聚类、分类领域引入迁移学习机制已进行了广泛的研究。如文献[17]将相同的词特征知识迁移到不同领域的聚类工作中，提出了同时利用文档和词特性进行的联合聚类方法。文献[18]通过在有紧密联系的任务之间进行有益知识的交叉学习，提出了共享子空间的多任务聚类方法。文献[19]提出的应用于图像聚类的异构迁移学习方法。文献[20]扩展了传统的潜在概率语义分析(PLSA)算法，提出了应用于交叉域文本分类的 TPLSA 算法。这些方法要能够顺利实施，需要有完整历史数据集作为支撑，这在一些特殊情况下是不可能满足的。如需要的历史数据集是一些敏感性的数据和有保密性要求的数据。

由于上述需求的存在，本文以模糊 C 均值(FCM)算法作为聚类算法的基本框架，在此框架的基础上通过引入迁移学习机制构造出一种具有隐私保护能力的迁移学习 KTCAPP 算法的目标函数。由于该算法通过对历史数据类中心和隶属度这两类知识的学习来指导当前数据量不足或被污染数据的聚类任务，从而能够使该算法获得比传统聚类算法更好的聚类结果。同时由于只需提供历史数据类中心和隶属度而不用完整的历史数据就可进行算法执行，故对历史数据具有很好的隐私保护作用，使该算法具备良好的应用价值。

2 经典的 FCM 算法

模糊 C 均值(FCM)聚类算法是 1973 年由 DUMA 首先提出，后由 BEZDEK 在 1981 年将它进行了进一步推广的一种聚类方法。FCM 算法因具有理论简单、操作方便、局部搜索能力强、收敛速度快等优点，是至今为止在模式识别、机器学习、数据挖掘等领域最常使用的聚类方法之一[7−11]。FCM 算法的目标函数一般表示为

$$\min J_{\mathrm{FCM}}(\boldsymbol{U},\boldsymbol{X},\boldsymbol{V})=\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}^{m}\left\|\boldsymbol{x}_i-\boldsymbol{v}_j\right\|^2,$$

$$\boldsymbol{\mu}_{ij}\in[0,1],\quad \sum_{j=1}^{C}\boldsymbol{\mu}_{ij}=1,\quad 1\le i\le N \tag{1}$$

其中 $\boldsymbol{\mu}_{ij}$ 表示第 i 个样本点属于第 j 类的隶属度，m 表示模糊度指数，m 必须满足 $m>1$，$\boldsymbol{x}_i$ 表示第 i 个样本点，$\boldsymbol{v}_j$ 表示第 j 类的中心点。模糊 C 均值算法是通过求解目标函数最小值而达到聚类目的的一种聚类方法。通过采用拉格朗日条件极值的优化理论获得式(2)，式(3)所示的最优聚类中心 $\boldsymbol{V}$ 以及隶属度 $\boldsymbol{U}$ 的迭代公式：

$$\boldsymbol{v}_j=\sum_{i=1}^{N}\boldsymbol{\mu}_{ij}^{m}\boldsymbol{x}_i\Big/\sum_{i=1}^{N}\boldsymbol{\mu}_{ij}^{m} \tag{2}$$

$$\boldsymbol{\mu}_{ij}=\left[\sum_{k=1}^{C}\left(\frac{||\boldsymbol{x}_i-\boldsymbol{v}_j||^2}{||\boldsymbol{x}_i-\boldsymbol{v}_k||^2}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{3}$$

FCM 算法的具体执行步骤如下：

步骤 1 设置迭代结束阈值 ε，最大迭代次数 iter_max，初始化隶属度矩阵 $\boldsymbol{U}$ 和迭代计数器 t=0；

步骤 2 根据式(2)计算新的类中心；

步骤 3 根据式(3)计算新的隶属度矩阵；

步骤 4 t=t+1；直至 $\left\|\boldsymbol{U}^{(t)}-\boldsymbol{U}^{(t-1)}\right\|<\varepsilon$ 或 $t>$ iter_max。

FCM 算法在迭代优化结束后，将得到的隶属度矩阵 $\boldsymbol{U}$ 去模糊化后即可得到每个样本 $\boldsymbol{x}_i$ 所对应的类别。

FCM 算法在样本数据量充足的情况下，通过以上迭代过程即可获得较理想的每个类中心，从而能够根据这些类中心获得每个样本点对各个类的隶属度。但当样本数据量不充足或被污染的情况下，根据该 FCM 算法获得的类中心将有严重偏离实际类中心的风险。有时甚至会导致聚类失败。所以在当前样本数据量不足或被污染的场景下寻找新的有效聚类方法有其必要性和实用性。

3 具有隐私保护功能的知识迁移聚类算法

根据迁移学习理论，历史场景与目标场景具有一定的相似性，那么对目标场景来说充分利用历史场景的有益知识必将对目标场景具有很好的指导作用。在模糊聚类算法中起关键作用的是各个类别的类中心和样本点对各个类别的隶属度。故利用好历史场景所导出的历史类中心和历史隶属度将对当前聚类任务起到积极作用。

根据以上分析我们提出了样本点与历史类中心点距离和极小规则。该规则的提出是基于历史场景和目标场景具有一定的相似性，那么当前目标场景中的样本点与历史类中心的距离和也必定达到最小，通过样本点与历史类中心点距离和极小规则可以使最终的聚类结果满足此约束。该规则用公式表示如式(4)：

$$\min J_{\mathrm{KT1}}(\boldsymbol{X},\boldsymbol{U},\widetilde{\boldsymbol{V}})=\lambda_1\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}^{2}\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_j\right\|^2,\ \lambda_1\ge 0 \tag{4}$$

其中 $\boldsymbol{x}_i$ 表示第 i 个样本点，$\tilde{\boldsymbol{v}}_j$ 表示第 j 类的历史类中心，$\boldsymbol{\mu}_{ij}$ 表示当前样本隶属度矩阵，N 表示样本点的数目，C 表示聚类数，λ_1 为样本点与历史类中心点距离和极小规则的平衡因子。

样本点与历史类中心点距离和极小规则的应用主要是迁移了历史类中心的知识，同样我们也可对历史隶属度的知识进行迁移。本文提出隶属度变化极小规则。该规则的提出同样是基于历史场景和目标场景具有一定的相似性，那么当前样本点的历史隶属度和当前隶属度应该变化尽量的小。通过隶属度变化极小规则的引入可以使最终的聚类结果满足此约束。该规则用公式表示为

$$\min J_{\mathrm{KT2}}\left(\boldsymbol{U},\widetilde{\boldsymbol{U}}\right)=\lambda_2\sum_{i=1}^{N}\sum_{j=1}^{C}\left(\boldsymbol{\mu}_{ij}-\widetilde{\boldsymbol{\mu}}_{ij}\right)^2,\ \ \lambda_2\geq 0 \tag{5}$$

其中 $\boldsymbol{\mu}_{ij}$ 表示当前样本隶属度矩阵，$\widetilde{\boldsymbol{\mu}}_{ij}$ 为当前样本点的历史隶属度矩阵，λ_2 为隶属度变化极小规则的平衡因子。

根据以上分析，针对模糊 C 均值算法没有知识迁移机制的不足，本文在模糊 C 均值算法的基础上融入了历史类中心和历史隶属度的双重知识迁移学习机制提出具有隐私保护功能的知识迁移聚类算法即 KTCAPP 算法。该算法的原理图如图 1 所示。

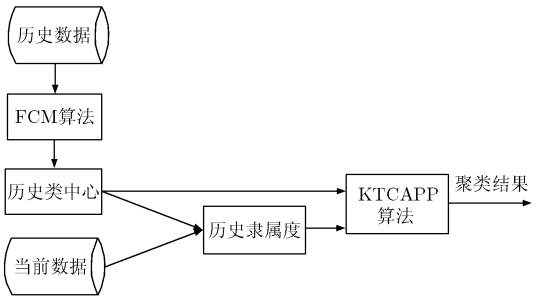

图1　KTCAPP算法原理图

图 1 中所涉及的历史类中心是根据历史数据利用经典的 FCM 算法所获得；历史隶属度是当前数据相对于历史类中心所对应的隶属度，它可利用式(3)和历史类中心获得。

KTCAPP 算法对应的目标函数为

$$\begin{aligned}&J_{\mathrm{PPKTFCM}}(\boldsymbol{X},\boldsymbol{V},\widetilde{\boldsymbol{V}},\boldsymbol{U},\widetilde{\boldsymbol{U}})\\&=J_{\mathrm{FCM}}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{U})+J_{\mathrm{KT1}}(\boldsymbol{X},\boldsymbol{U},\widetilde{\boldsymbol{V}})+J_{\mathrm{KT2}}(\boldsymbol{U},\widetilde{\boldsymbol{U}}),\\&\boldsymbol{\mu}_{ij}\in[0,1],\quad\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}=1,\quad 1\leq i\leq N\end{aligned} \tag{6a}$$

其中

$$J_{\mathrm{FCM}}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{U})=\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}^2\left\|\boldsymbol{x}_i-\boldsymbol{v}_j\right\|^2 \tag{6b}$$

$$J_{\mathrm{KT1}}(\boldsymbol{X},\boldsymbol{U},\widetilde{\boldsymbol{V}})=\lambda_1\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}^2\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_j\right\|^2 \tag{6c}$$

$$J_{\mathrm{KT2}}(\boldsymbol{U},\widetilde{\boldsymbol{U}})=\lambda_2\sum_{i=1}^{N}\sum_{j=1}^{C}\left(\boldsymbol{\mu}_{ij}-\widetilde{\boldsymbol{\mu}}_{ij}\right)^2 \tag{6d}$$

根据式(6a)可以看出，加入知识迁移机制的 KTCAPP 算法从本质上来说由 3 部分组成。第 1 部分 $J_{\mathrm{FCM}}(\boldsymbol{X},\boldsymbol{V},\boldsymbol{U})$ 为经典的 FCM 算法，对当前样本使用 FCM 算法进行聚类，第 2 部分 $J_{\mathrm{KT1}}(\boldsymbol{X},\boldsymbol{U},\widetilde{\boldsymbol{V}})$ 为利用历史类中心对当前聚类进行指导项，通过该项实现了历史类中心知识的迁移学习。第 3 部分 $J_{\mathrm{KT2}}(\boldsymbol{U},\widetilde{\boldsymbol{U}})$ 为利用历史隶属度对当前聚类任务进行指导项，通过该项实现了历史隶属度知识的迁移学习。在式(6b)~式(6d)中 $\boldsymbol{\mu}_{ij}$ 表示当前样本的隶属度矩阵，$\widetilde{\boldsymbol{\mu}}_{ij}$ 为历史隶属度矩阵，$\boldsymbol{x}_i$ 为当前样本点，$\boldsymbol{v}_j$ 为第 j 类的聚类中心，$\tilde{\boldsymbol{v}}_j$ 为第 j 类的历史聚类中心。λ_1 为历史类中心的平衡因子，λ_2 为历史隶属度的平衡因子。

根据式(6a)可以发现，当参数 $\lambda_1=0$ 且 $\lambda_2=0$ 时，KTCAPP 算法实际上就退化为经典的 FCM 算法。KTCAPP 算法本质上是在 FCM 的基础上通过调控 $J_{\mathrm{KT1}}(\boldsymbol{X},\boldsymbol{U},\widetilde{\boldsymbol{V}})$ 和 $J_{\mathrm{KT2}}(\boldsymbol{U},\widetilde{\boldsymbol{U}})$ 两部分的权重来实现知识的迁移达到对当前聚类任务的引导学习，从而改善因数据量不足或数据所含信息不足而直接采用 FCM 算法聚类效果不理想的缺陷。

3.1 算法优化

为了求得式(6a)在有约束条件下目标函数的最小值，可以利用拉格朗日乘子法进行优化求解，其拉格朗日函数如式(7)：

$$\begin{aligned}L(\boldsymbol{U},\boldsymbol{V},\boldsymbol{\beta})=&\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}^2\left\|\boldsymbol{x}_i-\boldsymbol{v}_j\right\|^2\\&+\lambda_1\sum_{i=1}^{N}\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}^2\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_j\right\|^2\\&+\lambda_2\sum_{i=1}^{N}\sum_{j=1}^{C}\left(\boldsymbol{\mu}_{ij}-\widetilde{\boldsymbol{\mu}}_{ij}\right)^2\\&+\sum_{i=1}^{N}\beta_i\left(1-\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}\right)\end{aligned} \tag{7}$$

式(7)中的 β_i 为 Lagrange 乘子。

通过 $\partial L/\partial\boldsymbol{\mu}_{ij}=0$ 得

$$\boldsymbol{\mu}_{ij}=\frac{1}{2}\frac{2\lambda_2\widetilde{\boldsymbol{\mu}}_{ij}+\beta_i}{\left\|\boldsymbol{x}_i-\boldsymbol{v}_j\right\|^2+\lambda_1\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_j\right\|^2+\lambda_2} \tag{8}$$

根据约束条件

$$\sum_{j=1}^{C}\boldsymbol{\mu}_{ij}=1 \tag{9}$$

将式(8)代入式(9)可得

$$\beta_i=2\frac{1-\sum_{k=1}^{C}\dfrac{\lambda_2\widetilde{\boldsymbol{\mu}}_{ik}}{\left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|^2+\lambda_1\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_k\right\|^2+\lambda_2}}{\sum_{k=1}^{C}\dfrac{1}{\left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|^2+\lambda_1\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_k\right\|^2+\lambda_2}} \tag{10}$$

将式(10)代入式(8)可得隶属度的迭代公式如式(11)：

$$\boldsymbol{\mu}_{ij}=\frac{1-\sum_{k=1}^{C}\dfrac{\lambda_2(\tilde{\boldsymbol{\mu}}_{ik}-\tilde{\boldsymbol{\mu}}_{ij})}{\left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|^2+\lambda_1\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_k\right\|^2+\lambda_2}}{\sum_{k=1}^{C}\dfrac{\left\|\boldsymbol{x}_i-\boldsymbol{v}_j\right\|^2+\lambda_1\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_j\right\|^2+\lambda_2}{\left\|\boldsymbol{x}_i-\boldsymbol{v}_k\right\|^2+\lambda_1\left\|\boldsymbol{x}_i-\tilde{\boldsymbol{v}}_k\right\|^2+\lambda_2}} \tag{11}$$

通过$\partial L/\partial \boldsymbol{v}_j=0$得中心点迭代公式为

$$\boldsymbol{v}_j=\sum_{i=1}^{N}\boldsymbol{\mu}_{ij}^2\boldsymbol{x}_i\Big/\sum_{i=1}^{N}\boldsymbol{\mu}_{ij}^2 \tag{12}$$

根据迭代式(11)和式(12)，最终可求得最优的当前数据的类中心和隶属度，再经过对隶属度的去模糊化处理后，可获得最终的聚类结果。

3.2 算法步骤

根据上述对 KTCAPP 算法的推导过程所得到的迭代公式，给出具体的算法步骤如下：

步骤 1　给出当前数据集$\boldsymbol{x}$，迭代结束阈值ε，最大迭代次数 iter_max，历史类中心$\tilde{\boldsymbol{v}}_j$，平衡因子λ_1,λ_2，初始化隶属度矩阵$\boldsymbol{u}$和迭代计数器 t=0；

步骤 2　根据式(3)计算历史隶属度$\tilde{\boldsymbol{\mu}}_{ij}$；

步骤 3　根据式(12)计算新的类中心；

步骤 4　根据式(11)计算新的隶属度矩阵；

步骤 5　置$t=t+1$，根据式(6a)计算当前目标函数值$J^{(t)}$；

直至$\left\|J^{(t)}-J^{(t-1)}\right\|<\varepsilon$或 $t>$ iter_max。

4　实验及结果分析

4.1 实验设置

本实验使用的硬件环境：Intel i5 2.27 GHz 4 G RAM，操作系统：Windows 7 64 位，软件环境：MATLAB R2012b。

为验证本文所提出 KTCAPP 算法的有效性，本节将基于一组模拟数据集和两种常见真实数据集(20Newsgroups[17]和 KDDCups[21])进行实验。并且，还选择了与 KTCAPP 算法相关的另外 5 种算法(基于多任务共享子空间的 LSSMTC 聚类算法[18]，基于组合 K 均值的 CombKM 算法[18]，基于特征和样本协同的 Co-clustering 聚类算法[22]以及基于自学习迁移的 STC 聚类算法[23])进行聚类性能对比。

本实验采用归一化互信息(Normalized Mutual Information, NMI[24])和芮氏指标(Rand Index, RI[25])两种评价指标对实验结果进行评价。该两种评价指标的取值范围都为[0,1]，取值越大所对应算法的聚类性能越好。

本文算法涉及两个平衡因子λ_1和λ_2的取值问题。对于聚类算法中此类参数值的选择问题，现在惯用的方法是采用网格搜索来进行遍历寻优。本文也同样采用该方法，并设置λ_1和λ_2的 3 个寻优区间为：[0.1,1.0]，步距为 0.1; [2,10]，步距为 1; [20,100]，步距为 10。对于其它 5 种对比算法中的参数则采用其文献推荐的设置。

实验结果所列数据都是在运行 10 次后求均值所得。

4.2 模拟数据集实验及结果分析

本文构造如下 4 种类型的模拟数据集：

(1)历史数据集 $H1$, $H2$：　采用高斯概率分布模型函数生成 3 类共 1500 个数据的历史数据集 $H1$，每类 500 个数据。该 3 类数据样本满足以下分布：第 1 类均值$\boldsymbol{m}_1=\begin{bmatrix}2\\5\end{bmatrix}$，方差$\boldsymbol{S}_1=\begin{bmatrix}8&0\\0&8\end{bmatrix}$；第 2 类均值$\boldsymbol{m}_2=\begin{bmatrix}8\\16\end{bmatrix}$，方差$\boldsymbol{S}_2=\begin{bmatrix}16&0\\0&9\end{bmatrix}$；第 3 类均值$\boldsymbol{m}_3=\begin{bmatrix}6\\28\end{bmatrix}$，方差$\boldsymbol{S}_3=\begin{bmatrix}25&0\\0&15\end{bmatrix}$。产生与历史数据集 $H1$ 具有相同的高斯概率分布而数据量更大的历史数据集 $H2$。数据集 $H2$ 包含 3 类，每类 5000 个数据，共 15000 个数据。该两个历史数据集样本量充足，而且能通过该样本获取对目标数据集聚类具有指导作用的有益知识。

(2)样本量不足的目标集 $D1$：　采用高斯概率分布模型函数生成待测试的目标数据集 $D1$, $D1$ 的类别数也为 3 类。其数据满足如下分布：第 1 类均值$\boldsymbol{m}_1'=\begin{bmatrix}2.5\\5.0\end{bmatrix}$，方差$\boldsymbol{S}_1'=\begin{bmatrix}8&0\\0&8\end{bmatrix}$；第 2 类均值$\boldsymbol{m}_2'=\begin{bmatrix}8\\15\end{bmatrix}$，方差$\boldsymbol{S}_2'=\begin{bmatrix}16&0\\0&9\end{bmatrix}$；第 3 类均值$\boldsymbol{m}_3'=\begin{bmatrix}6\\27\end{bmatrix}$，方差$\boldsymbol{S}_3'=\begin{bmatrix}25&0\\0&15\end{bmatrix}$。$D1$ 的数据样本总量为 90 个，每类包含 30 个数据。$D1$ 数据样本量只为历史数据集 $H1$ 的 10%，用于代表样本量不足的场景。并且数据集 $D1$ 与数据集 $H1$ 的均值有略微不同，方差相同，体现迁移学习中的被学习的历史数据集与目标数据集之间既存在着很大的相似性，同时又存在着一定的差别的情况。

(3)样本量充足且无污染的目标集 $D2$：　再次采用高斯概率分布模型函数生成待测试的目标数据集 $D2$，其数据分布与 $D1$ 数据集相同。$D2$ 数据集包含 3 类，每类 150 个，共计 450 个数据，占历史数据集 $H1$ 的 50%，用于代表目标数据量较充分的场景。

(4)样本量充足但被污染的目标集 $D3$：　通过在 $D2$ 数据集上加入均值为 0、方差为 2 的高斯噪声形成样本量充足但受噪声污染的目标数据集 $D3$。

上述生成的 4 种不同模拟数据集的分布如图 2 所示。在以上构建的历史数据集和不同目标数据集的场景下，分别运行本文的 KTCAPP 算法和相关的 5 种对比算法，得到表 1 所列的实验结果。

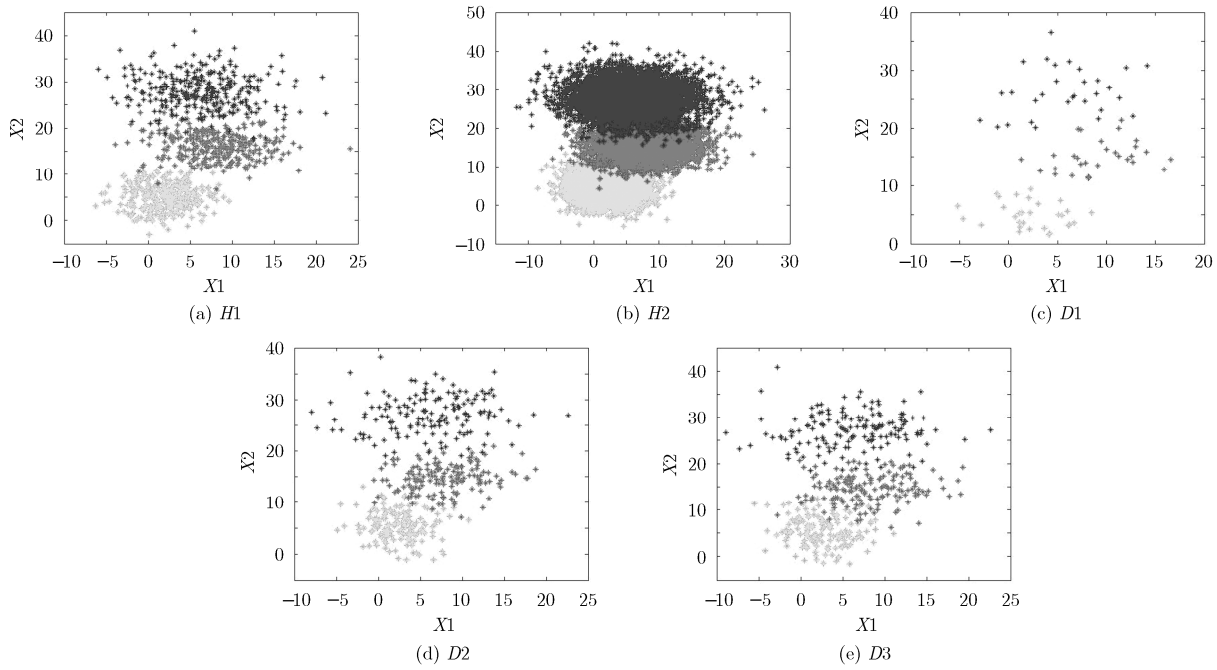

图 2 模拟数据集分布图

表 1 模拟数据集上 6 种聚类算法的性能对比

场景	评价指标	LSSMTC	CombKM	STC	Co-clustering	FCM	KTCAPP
*H*1-*D*1	NMI-mean	0.6393	0.7594	0.7580	0.7372	0.7889	**0.8056**
	NMI-std	1.1703E-16	1.1703E-16	0	1.1703E-16	1.1703E-16	1.1703E-16
	RI-mean	0.8624	0.9024	0.9112	0.9014	0.9151	**0.9278**
	RI-std	0	1.1703E-16	0	2.3406E-16	0	0
*H*1-*D*2	NMI-mean	0.8235	0.8335	0.8506	0.8557	0.8280	**0.8592**
	NMI-std	0	0	0	1.1703E-16	9.0649E-17	0
	RI-mean	0.9373	0.9380	0.9506	0.9488	0.9354	**0.9513**
	RI-std	0	1.1703E-16	0	1.1703E-16	2.3406E-16	0
*H*1-*D*3	NMI-mean	0.7280	0.7737	0.8100	0.7718	0.7737	**0.8160**
	NMI-std	1.1703E-16	1.1703E-16	0	1.1703E-16	1.1703E-16	1.1703E-16
	RI-mean	0.8966	0.9147	0.9350	0.9140	0.9147	**0.9374**
	RI-std	1.1703E-16	1.1703E-16	0	1.1703E-16	1.1703E-16	1.1703E-16
*H*2-*D*1	NMI-mean	0.6892	0.7594	0.8041	0.7372	0.7889	**0.8056**
	NMI-std	1.1703E-16	1.1703E-16	0	1.1703E-16	1.1703E-16	1.1703E-16
	RI-mean	0.8790	0.9024	0.9241	0.9014	0.9151	**0.9278**
	RI-std	0	1.1703E-16	0	2.3406E-16	0	0
*H*2-*D*2	NMI-mean	0.8188	0.8182	0.8356	0.8557	0.8280	**0.8592**
	NMI-std	0	1.6131E-16	0	1.1703E-16	9.0649E-17	0
	RI-mean	0.9352	0.9301	0.9448	0.9488	0.9354	**0.9513**
	RI-std	0	1.1703E-16	0	1.1703E-16	2.3406E-16	0
*H*2-*D*3	NMI-mean	0.7242	0.7777	0.7904	0.7718	0.7737	**0.8160**
	NMI-std	1.1703E-16	1.1703E-16	0	1.1703E-16	1.1703E-16	1.1703E-16
	RI-mean	0.8957	0.9147	0.9266	0.9140	0.9147	**0.9374**
	RI-std	1.1703E-16	2.3406E-16	0	1.1703E-16	1.1703E-16	1.1703E-16

通过表 1 的数据可以获得如下结论：

(1)由源域是历史数据集 $H1$，目标域是样本量不足的 $D1$ 数据集组成的场景 $H1$-$D1$ 上的实验结果可以看出，当聚类样本量不足的情况下直接使用基于无任何学习机制的 FCM 算法，其产生的聚类结果明显要差于使用迁移知识的 KTCAPP 算法。产生此结果的原因是：在聚类样本量不足的情况下直接使用 FCM 算法得到的聚类中心将会严重偏离真正的聚类中心，正是因为存在着这样的偏离使得 FCM 算法的聚类效果要差于 KTCAPP 算法的聚类效果。而 KTCAPP 算法通过对历史类中心和隶属度的学习，使得由于数据量不足而导致聚类中心产生偏移的问题得到有效抑制，从而提高了最终的聚类性能。同时，从其它 4 个算法的聚类结果可以看出，由于自身算法的限制，它们在数据样本量不足场景下的聚类效果不理想。

(2)由源域是历史数据集 $H1$，目标域是样本量充足且无污染的 $D2$ 数据集组成的场景 $H1$-$D2$ 上的实验结果可以看出，由于 $D2$ 的样本数据量较为充分，因此 6 个算法的实验结果都比较理想。由于 KTCAPP 算法有效学习了历史类中心和隶属度知识的原因，使得 KTCAPP 算法跟其它算法相比，还是取得了最好的聚类结果。

(3)由源域是历史数据集 $H1$，目标域是样本量充足但被污染的 $D3$ 数据集组成的场景 $H1$-$D3$ 上的实验结果可以看出，在样本数据被污染的情况下，KTCAPP 算法和 STC 算法的结果比较理想，而其它 4 种算法性能明显变差。产生此现象的主要原因是：由于 $D3$ 数据集受到了污染，数据也因此产生了严重失真，使得数据的类中心产生偏移，数据类别之间的边界变得模糊，最终造成聚类结果不理想。而 KTCAPP 算法在进行聚类任务时通过调整两个平衡因子 λ_1 和 λ_2 的值，即调整历史类中心和历史隶属度这两类指导性的知识在整个聚类任务中权重，使得历史知识得到充分利用，最终使得在数据被污染的场景下该算法也能取得较理想的聚类结果。

(4)当源域选择样本量更大的历史数据集 $H2$，目标域使用数据集 $D1$, $D2$, $D3$ 组成的场景 $H2$-$D1$, $H2$-$D2$, $H2$-$D3$ 上的实验结果可以看出，FCM 算法在目标数据集相同，历史数据集不同的情况下产生了相同的聚类结果。这是因为 FCM 算法仅通过目标域进行聚类，与源域无关，所以产生的是相同的结果。KTCAPP 算法在目标数据集相同，源域不同的场景下同样产生相同的聚类结果。这主要是因为 KTCAPP 算法仅利用了由源域产生的历史类中心和历史隶属度，与样本量的大小无关，而历史数据集 $H1$ 和 $H2$ 使用的是同分布，由此得出的历史类中心和历史隶属度相同，所以产生了相同的聚类结果。同样，因为 KTCAPP 算法的有效利用迁移知识的机制，产生了比其它算法更好的聚类效果。

4.3 文本数据集实验及结果分析

本文选取常用的 20 Newsgroups(20NG)数据集来构造两个文本迁移场景：“comp VS sci”和“rec VS talk”。不同迁移场景所使用的数据集的来源见表 2 所示。表 3 列出了不同迁移场景下的数据集的样本量、数据维数和包含的类别数。在使用 20NG 数据之前，通过 BOW[26]工具对其原始数据进行了降维预处理。6 种不同聚类算法在该文本迁移场景上的实验结果如表 4 所示。

表 2 文本迁移场景的数据集来源

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

表 3 文本迁移场景数据集构成

迁移场景	数据集	样本个数	维数	类别
comp VS sci	历史数据集	1500	350	2
	目标数据集	150	350	
rec VS talk	历史数据集	1500	350	2
	目标数据集	150	350	

通过表 4 的数据可以看出：

(1)KTCAPP 算法和 STC 算法在“comp VS sci” 迁移场景上聚类效果优势明显。分析其原因，两者都是因为其算法中使用了迁移学习的机制，通过对有用历史知识的学习，大大提高了聚类效果。

(2)KTCAPP 算法和 STC 算法在“rec VS talk”迁移场景上聚类结果同样优势明显，其中 KTCAPP 算法聚类效果最好。LSSMTC 算法、CombKM 算法和 Co-clustering 算法虽然也引入各种学习机制，但聚类效果不明显，分析其原因，KTCAPP 算法和 STC 算法的知识学习机制比上述这些算法在该场景下更有效。

(3)通过观察表 4 的数据可以进一步发现 KTCAPP 算法的聚类性能在两种迁移场景下始终都高于经典的 FCM 算法，这是由于 KTCAPP 的迁移学习机制可通过调节平衡因子，从而达到对有用历史知识的有效利用，确保 KTCAPP 算法的聚类效果比 FCM 算法更好。

表 4 文本迁移场景上 6 种聚类算法性能对比

数据集	评价指标	LSSMTC	CombKM	STC	Co-clustering	FCM	KTCAPP
comp VS sci	NMI-mean	0.2127	0.1519	0.6863	0.2127	0.2449	**0.7457**
	NMI-std	1.1703E-16	0.1961	0	0.1323	2.9257E-17	0.0090
	RI-mean	0.5400	0.5568	0.8492	0.5400	0.5571	**0.8960**
	RI-std	1.1703E-16	0.0777	0	0.0556	1.1703E-16	0.0050
Rec VS talk	NMI-mean	0.0279	0.0862	0.8071	0.0028	0.2310	**0.9015**
	NMI-std	1.1703E-16	0.0412	0	0.0088	2.9257E-17	0.0213
	RI-mean	0.4967	0.5021	0.9370	0.5027	0.6063	**0.9696**
	RI-std	1.1703E-16	0.0045	0	0.0021	1.1703E-16	0.0087

4.4 入侵检测数据集实验及结果分析

本文选择 KDDCup99 网络入侵检测数据集构造另一个迁移场景。KDDCup99 数据集中的网络连接数据是从一个模拟美国空军的局域网收集而来。其中包含了带有类标签的训练数据和不带类标签的测试数据。测试数据和训练数据有着不同的概率分布，符合迁移场景的历史数据集和目标数据集之间在特征空间总体上相似，但存在着一定的差异性的条件。本文选取 KDDCup99 数据集中 5 种常见类型(Normal, smurf, Neptune, satan, ipsweep)的数据作为历史数据集和目标数据集的类别。其中历史数据集来自训练样本，目标数据集来自测试样本。目标数据集的样本量只有历史数据集样本量的 10%，体现了目标数据集样本量不足的场景。选择每个网络连接记录中的 32 个连续型的特征属性作为样本的维度。历史数据集和目标数据集的构成如表 5 所示。

表 5 入侵检测迁移场景的数据集构成

数据集	样本个数	维数	类别
历史数据集	2000	32	5
目标数据集	200	32	

本文算法和 5 个对比算法在此网络入侵检测迁移场景上执行的结果如表 6 所示。

从表 6 的数据可以看出，KTCAPP 算法在两大性能指标上明显要高于其它 5 种算法，这是由于 KTCAPP 算法从历史数据集中学习了有益的知识，并对目标数据集的聚类形成有效指导，从而使得该算法与其它 5 种算法相比具有更好的聚类性能。这也进一步验证了对已有知识的有效学习会对当前聚类任务的有效执行起到积极作用。

5　结束语

本文针对传统聚类算法在数据样本量不足和数据样本受污染的场景下往往聚类结果不理想甚至失效的问题，在经典 FCM 算法的基础上引入了迁移学习机制而提出了具有隐私保护功能的知识迁移聚类算法，即 KTCAPP 算法。该算法通过对历史相关数据中总结出的类中心和隶属度的有益知识的学习来指导当前聚类任务。该算法在模拟数据集、文本数据集以及入侵检测数据集上的实验结果以及与 5 种相关算法所进行的性能比对与分析均显示出了本文算法的有效性和优越性。同时，由于本文方法无需完整历史样本的支持而只需历史类中心和历史隶属度即可指导当前的聚类任务进行，使得本文方法对历史数据具有隐私保护能力及更好的实用价值。本文的创新点主要有两个，一是通过有效学习历史数据的类中心和隶属度两种迁移知识，改进了经典 FCM 算法在数据不足和数据被污染环境下的聚类性能不佳问题；二是本文所提算法在迁移学习过程中对历史数据具有隐私保护能力。此外，本文

表 6 入侵检测迁移场景上 6 种聚类算法性能对比

评价指标	LSSMTC	CombKM	STC	Co-clustering	FCM	KTCAPP
NMI-mean	0.3819	0.3907	0.3970	0.0343	0.3823	**0.4462**
NMI-std	1.1703E-16	0.0309	0	0.1085	0	1.1703E-16
RI-mean	0.4864	0.4983	0.5070	0.3668	0.5040	**0.5484**
RI-std	1.1703E-16	0.0302	0	0.0477	0	5.8514E-17

算法也存在着缺陷，如模糊度指数 m 需取固定值 2。这限制了该算法的一般化。对于模糊度指数 m 取更一般化的值的情况，也是将来对该算法进一步研究的方向。

参 考 文 献

[1] FERRARI D G and CASTRO L N. Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods [J]. *Information Sciences*, 2015, 301(1): 181–194. doi: 10.1016/j.ins.2014.12.044.

[2] TZORTZIS G and LIKAS A. The minmax k-means clustering algorithm[J]. *Pattern Recognition*, 2014, 47(7): 2505–2516. doi: 10.1016/j.patcog.2014.01.015.

[3] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48–61. doi: 10.3724/SP.J.1001.2008.00048.
SUN J G, LIU J, and ZHAO L Y. Clustering algorithms research[J]. *Journal of Software*, 2008, 19(1): 48–61. doi: 10.3724/SP.J.1001.2008.00048.

[4] 邓赵红, 张江滨, 蒋亦樟, 等. 基于模糊子空间聚类的 0 阶 L2 型 TSK 模糊系统[J]. 电子与信息学报, 2015, 37(9): 2082–2088. doi: 10.11999/JEIT150074.
DENG Z H, ZHANG J B, JIANG Y Z, *et al*. Fuzzy subspace clustering based zero-order L2-norm TSK fuzzy system[J]. *Journal of Electronics & Information Technology*, 2015, 37(9): 2082–2088. doi: 10.11999/JEIT150074.

[5] POPAT S K and EMMANUEL M. Review and comparative study of clustering techniques[J]. *International Journal of Computer Science and Information Technologies*, 2014, 5(1): 805–812.

[6] BOUGUETTAYA A, YU Q, LIU X, *et al*. Efficient agglomerative hierarchical clustering[J]. *Expert Systems with Applications*, 2015, 42(5): 2785–2797. doi: 10.1016/j.eswa.2014.09.054.

[7] ZHU L, CHUNG F L, and WANG S T. Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions[J]. *IEEE Transactions on System, Man and Cybernetics*, 2009, 39(3): 578–591. doi: 10.1109/TSMCB.2008.2004818.

[8] DUNN J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. *Journal of Cybernetics*, 1973, 3(3): 32–57. doi: 10.1080/01969727308546046.

[9] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York, Plenum Press, 1981: 43–93.

[10] 赵凤, 刘汉强, 范九伦. 基于互补空间信息的多目标进化聚类图像分割[J]. 电子与信息学报, 2015, 37(3): 672–678. doi: 10.11999/JEIT140371.
ZHAO F, LIU H Q, and FAN J L. Multi-objective evolutionary clustering with complementary spatial information for image segmentation[J]. *Journal of Electronics & Information Technology*, 2015, 37(3): 672–678. doi: 10.11999/JEIT140371.

[11] 赵雪梅, 李玉, 赵泉华. 结合高斯回归模型和隐马尔可夫随机场的模糊聚类图像分割[J]. 电子与信息学报, 2014, 26(11): 2730–2736. doi: 10.3724/SP.J.1146.2013.01751.
ZHAO X M, LI Y, and ZHAO Q H. Image segmentation by fuzzy clustering algorithm combining hidden Markov random field and Gaussian regression model[J]. *Journal of Electronics & Information Technology*, 2014, 26(11): 2730–2736. doi: 10.3724/SP.J.1146.2013.01751.

[12] KIM Y H, SHIM K, KIM M S, *et al*. DBCURE-MR: An efficient density-based clustering algorithm for large data using MapReduce[J]. *Information Systems*, 2014, 42(1): 15–35. doi: 10.1016/j.is.2013.11.002.

[13] AGRAWAL A S and BOJEWWAR S. Comparative study of various clustering techniques[J]. *International Journal of Computer Science and Mobile Computing*, 2014, 3(10): 497–504.

[14] SHAO L, ZHU F, and LI X. Transfer learning for visual categorization: a survey[J]. *Neural Networks and Learning*, 2014, 26(5): 1019–1034. doi: 10.1109/TNNLS.2014.2330900.

[15] LU J, BEHBOOD V, HAO P, *et al*. Transfer learning using computational intelligence: A survey[J]. *Knowledge-based Systems*, 2015, 80(1): 14–23. doi: 10.1016/j.knosys.2015.01.010.

[16] LONG M S, WANG J M, DING G G, *et al*. Transfer learning with graph co-regularization[J]. *Knowledge and Data Engineering*, 2014, 26(7): 1805–1818. doi: 10.1109/TKDE.2013.97.

[17] DAI W Y, XUE G R, YANG Q, *et al*. Co-clustering based classification for out-of-domain document[C]. The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2007: 210–219. doi: 10.1145/1281192.1281218.

[18] GU Q and ZHOU J. Learning the shared subspace for multi-task clustering and transductive transfer classification[C]. The 2009 Ninth IEEE International Conference on Data Mining, IEEE, Washington DC, USA, 2009: 159–168. doi: 10.1109/ICDM.2009.32.

[19] YANG Q, CHEN Y Q, XUE G R, *et al*. Heterogeneous transfer learning for image clustering via the social web[C]. Proceeding of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, Suntec, Singapore, 2009: 1–9.

[20] XUE G R, DAI W Y, YANG Q, *et al*. Topic-bridged PLSA for cross-domain text classification[C]. Proceedings of the 31st Annual International ACM SIGIR Conference on

Research and Development in Information Retrieval, New York, NY, USA, ACM, 2008: 627–634. doi: 10.1145/1390334.1390441.

[21] MOHAMMAD K S and SHAMS N. Analysis of KDD CUP 99 dataset using clustering based data mining[J]. *International Journal of Database Theory and Application*, 2013, 6(5): 23–34.

[22] GU Q and ZHOU J. 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. doi: 10.1145/1557019.1557063.

[23] DAI W Y, YANG Q, XUE G R, *et al*. Self-taught clustering [C]. Proceeding of the 25th International Conference on Machine Learning, ACM, New York, NY, USA, 2008: 200–207. doi: 10.1145/1390156.1390182.

[24] JING L, NG K M, and HUANG 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. doi: 10.1109/TKDE.2007.1048.

[25] LIU J, MOHAMMED J, CARTER J, *et al*. Distance-based clustering of CGH data[J]. *Bioinformatics*, 2006, 22(16): 1971–1978. doi: 10.1093/bioinformatics/btl185.

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

陈爱国： 男，1975 年生，博士生，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，研究方向为人工智能和机器学习.