DOI:10.11992/tis.201603046
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160513.0957.034.html

一种基于少量标签的改进迁移模糊聚类

王跃,杨燕,王红军
(西南交通大学 信息科学与技术学院,四川 成都 610031)

摘 要:传统聚类算法难以利用已有的历史信息,尤其是数据被污染的情况下聚类结果不理想;半监督聚类常用于数据中有部分标签的情况。在源数据有少量标签的情况下,提出半监督混合 C 均值聚类算法(SS-FPCM);基于迁移学习框架,针对负迁移问题对算法进行修正,提出了防止负迁移的半监督迁移算法(TSS-FPCM);最后,为了充分借鉴源数据的信息,利用"代表点"来代替源数据类信息,融入算法中再次迁移得到改善的半监督迁移算法(ITSS-FPCM)。实验表明,3 个算法能够有效的利用源数据提高聚类性能。SS-FPCM 与 TSS-FPCM 可以利用源数据的少量标签数据,而 ITSS-FPCM 算法结合了标签数据与"代表点"两个有效信息,在数据信息匮乏、数据被污染的情况下得到较好的聚类结果。

关键词:聚类;迁移学习;半监督;可能性 C 均值;模糊 C 均值
中图分类号:TP301 **文献标志码**:A **文章编号**:1673-4785(2016)03-0310-08

中文引用格式:王跃,杨燕,王红军.一种基于少量标签的改进迁移模糊聚类[J]. 智能系统学报, 2016, 11(3): 310-317.
英文引用格式:WANG Yue, YANG Yan, WANG Hongjun.An improved transfer fuzzy clustering with few labels[J]. CAAI transactions on intelligent systems, 2016,11(3): 310-317.

An improved transfer fuzzy clustering with few labels

WANG Yue, YANG Yan, WANG Hongjun
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract: In the traditional clustering algorithm, it is difficult to utilize existing historical information, which tends to be less effective in cases in which the data is contaminated. The semi-supervised clustering algorithm is often used in such circumstances, wherein the target data has some labeled examples. For situations in which the source data has partially labeled samples, in this paper, we propose a semi-supervised fuzzy possibilistic C-means algorithm (SS-FPCM). Based on the transfer learning framework, we use a transfer semi-supervised fuzzy possibilistic C-means algorithm (TSS-FPCM) to avoid the negative transfer learning problem. Finally, in order to make full use of source data information, we use representative points to replace the source data class. Thus, we have developed an improved transfer semi-supervised fuzzy possibilistic C-means algorithm (ITSS-FPCM). The experimental results demonstrate that these three algorithms may be used to improve the clustering performance by using source data effectively, as compared with other clustering algorithms. Moreover, the SS-FPCM and TSS-FPCM algorithms exploit partially labeled data from the source, while the ITSS-FPCM algorithm combines the labeled data and "representative points," for cases having insufficient data information or contaminated data, and an excellent clustering result is attained.

Keywords: clustering; transfer learning; semi-supervised; possibilistic C-means; fuzzy C-means

传统的聚类算法在拥有大量数据的情况下能够在不同的场景下发挥各自的作用,当数据匮乏、噪声污染的情况,传统的聚类算法存在着不足。

近年来,迁移学习的成果逐渐丰富,研究表明,迁移学习能够有效地解决数据量不足、数据受污染和信息丢失等问题。文献[1]根据迁移学习中源领域和目标领域中是否含有标签,可以将迁移学习划

收稿日期:2016-03-19. **网络出版日期**:2016-05-13.
基金项目:国家自然科学基金项目(61170111, 61572407, 61134002);四川省科技支撑计划项目(2014SZ0207).
通信作者:杨燕. E-mail: yyang@swjtu.edu.cn.

分为 3 类:归纳迁移学习、直推式迁移学习和无监督迁移学习。现有的迁移学习在分类领域已有较多研究成果[2-10],而在聚类领域迁移学习理论和方法相对则要少很多。文献[11-12]在聚类领域利用了迁移学习的理论。

半监督聚类是半监督学习与聚类分析相结合的研究领域,文献[13]提出了不同情况下的半监督聚类算法,并取得了不错的效果。

文献[14]将经典的模糊 C 均值算法[15](FCM)与可能性 C 均值[16](PCM)算法进行改进,提出了模糊可能性聚类算法(FPCM)。本文探讨在源领域有少量标签的情况下,如何指导目标域进行聚类,提出半监督模糊可能性 C 均值聚类算法(SS-FPCM),并针对负迁移问题对算法进行改进,提出了防止负迁移的半监督迁移算法(TSS-FPCM),同时,用代表点代替源领域的数据进行数据迁移,得到改善的半监督迁移算法(ITSS-FPCM),并进行了实验验证。

1　相关算法介绍

1.1　PCM 聚类算法

PCM 聚类算法放松了传统 FCM 聚类算法中对于隶属度矩阵的约束,隶属度不再是对 1 的共享。对于给定数据集 $X=\{\boldsymbol{x}_k \mid k=1,2,\cdots,N\}$,$\boldsymbol{x}_k \in R^d$,包含 N 个样本,分成 C 个类别,$\boldsymbol{T}=\{t_{ik} \mid i=1,2,\cdots,C;k=1,2,\cdots,N\}$ 是可能划分矩阵,t_{ik} 表示第 k 个样本 x_k 属于第 i 类的可能性,聚类中心为 $V=\{\boldsymbol{v}_i \mid i=1,2,\cdots,C\}$,其中 $\boldsymbol{v}_i$ 表示第 i 个聚类中心。PCM 目标函数定义为

$$J=\sum_{i=1}^{C}\sum_{k=1}^{N} t_{ik}^m d_{ik}^2+\sum_{i=1}^{C}\eta_i\sum_{k=1}^{N}(1-t_{ik})^m d_{ik}^2 \quad (1)$$

式中:$t_{ik}\in[0,1]$,$0<\sum_{k=1}^{N} t_{ik}^m \leqslant N$,$m$ 为模糊指数,d_{ik}^2 和 η_i 的取值分别为式(2)和式(3),K 的取值一般取 $K=1$。

$$d_{ik}^2=\|\boldsymbol{x}_k-\boldsymbol{v}_i\|^2=(\boldsymbol{x}_k-\boldsymbol{v}_i)^{\mathrm{T}}(\boldsymbol{x}_k-\boldsymbol{v}_i) \quad (2)$$

$$\eta_i=K\frac{\sum_{k=1}^{N} t_{ik}^m d_{ik}^2}{\sum_{k=1}^{N} t_{ik}^m},K>0 \quad (3)$$

最小化目标函数可以得到可能性矩阵和聚类中心的迭代式(4)和式(5):

$$\boldsymbol{t}_{ik}=\frac{1}{1+\left(\frac{d_{ik}^2}{\eta_i}\right)^{\frac{1}{m-1}}},\forall i,k \quad (4)$$

$$\boldsymbol{v}_i=\frac{\sum_{k=1}^{N} t_{ik}^m x_k}{\sum_{k=1}^{N} t_{ik}^m},\forall i \quad (5)$$

1.2　PFCM 聚类算法

FPCM 是建立在 FCM 和 PCM 基础上的算法,它将两者结合在一起,FPCM 的目标函数定义为

$$J=\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik}^m+t_{ik}^\eta)\,d_{ik}^2 \quad (6)$$

式中:$m>1$,$\eta>1$,$0\leqslant_{ik}$,$t_{ik}\leqslant 1$,约束条件为

$$\sum_{k=1}^{N} t_{ik}=1,\forall i \quad (7)$$

$$\sum_{i=1}^{C} u_{ik}=1,\forall k \quad (8)$$

通过最小化目标函数,可以得到以下迭代优化公式:

$$\boldsymbol{u}_{ik}=\left[\sum_{j=1}^{C}\left(\frac{d_{ik}^2}{d_{ij}^2}\right)^{\frac{1}{m-1}}\right]^{-1},\forall i,k \quad (9)$$

$$\boldsymbol{t}_{ik}=\left[\sum_{j=1}^{N}\left(\frac{d_{ik}^2}{d_{ij}^2}\right)^{\frac{1}{\eta-1}}\right]^{-1},\forall i,k \quad (10)$$

$$\boldsymbol{v}_i=\frac{\sum_{k=1}^{N}(u_{ik}^m+t_{ik}^\eta)\,\boldsymbol{x}_k}{\sum_{k=1}^{N}(u_{ik}^m+t_{ik}^\eta)},\forall i \quad (11)$$

1.3　半监督聚类算法

对于一些有着一部分标签的数据集,在文献[17]中,Pedrycz 提出了基于部分标签的模糊聚类算法(SS-FCM),算法的核心思想是利用现有的分类信息,并把它作为优化程序的一部分。

为了区分标记数据与未标记数据,引入向量矩阵 $\boldsymbol{B}=\{b_k \mid k=1,2,\cdots,N\}$,如果 $\boldsymbol{x}_k$ 是已知标签样本 $b_k=1$,否则 $b_k=0$。并且记类别属性 $\boldsymbol{F}=\{f_{ik} \mid i=1,2,\cdots,C;k=1,2,\cdots,N\}$,如果 $\boldsymbol{x}_k$ 属于第 i 类,那么 $f_{ik}=1$;否则 $f_{ik}=0$。在引入 $\boldsymbol{B}$ 和 $\boldsymbol{F}$ 后,Pedrycz 将模糊参数 m 取值为 2,其目标函数为

$$J=\sum_{i=1}^{C}\sum_{k=1}^{N} u_{ik}^2 d_{ik}^2+\alpha\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik}-f_{ik}b_k)^2 d_{ik}^2 \quad (12)$$

2　半监督迁移模糊聚类算法

2.1　半监督模糊可能性 C 均值聚类算法

对半监督 FCM 算法进行研究可以发现,上文中的 $\boldsymbol{B}$ 和 $\boldsymbol{F}$ 的功能相似,保留下 $\boldsymbol{F}$ 并对 FPCM 的目标函数做如下改进:

$$J=\sum_{i=1}^{C}\sum_{k=1}^{N}(\alpha u_{ik}^2+\beta t_{ik}^2)\,d_{ik}^2+\omega\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik}-f_{ik})^2 d_{ik}^2 \tag{13}$$

s.t. $\alpha\geqslant 0,\ \beta\geqslant 0,\ \omega>0,\ 0\leqslant u_{ik},\ t_{ik}\leqslant 1$

最小化目标函数,可以得到迭代表达式:

$$\boldsymbol{t}_{ik}=\left[\sum_{j=1}^{N}\left(\frac{d_{ik}^2}{d_{ij}^2}\right)\right]^{-1},\ \forall i,k \tag{14}$$

$$\boldsymbol{u}_{ik}=\frac{1}{\alpha+\omega}\frac{\alpha+\omega\left(1-\sum_{j=1}^{C}f_{jk}\right)}{\sum_{j=1}^{C}\frac{d_{ik}^2}{d_{jk}^2}}+\omega f_{ik},\ \forall i,k \tag{15}$$

$$\boldsymbol{v}_i=\frac{\sum_{k=1}^{N}[\alpha u_{ik}^2+\beta t_{ik}^2+\omega(u_{ik}-f_{ik})^2]\,x_k}{\sum_{k=1}^{N}[\alpha u_{ik}^2+\beta t_{ik}^2+\omega(u_{ik}-f_{ik})^2]},\ \forall i \tag{16}$$

通过不断迭代优化隶属度矩阵最终获得我们需要的划分。改进的半监督模糊可能性 C 均值算法(SS-FPCM)能够通过 α、β 控制 FPCM 中 FCM 和 PCM 的权重,通过参数 ω 的变化控制已知标签在算法中所占的比重。

2.2 历史标签数据的迁移

迁移学习可以将历史场景(也叫源数据)中获取需要的数据或者信息,用于指导当前场景(又成为目标数据),当历史场景的信息与当前场景的相关性足够大时,可以从中得到潜藏的信息。在当历史场景没有任何指导信的数据(无任何标签信息)时,文献[11-12]针对这种情况分别做出了自己的研究。

当源数据有少量的标签时候,可以很直观地想到,将这些数据提取出来,加入到当前场景,一起进行聚类,以期待能够指导当前场景。前面提到了半监督 FPCM 聚类算法能够有效利用标签进行聚类,便可以直接引用式(13)的目标函数。但是,在迁移学习中负迁移是难以避免的一个问题,如果历史场景与当前场景相关性并不大。那么历史数据的标签很可能对当前场景产生不良影响,造成负迁移现象。针对这个问题,对式(13)进行改造,提出避免负迁移的半监督迁移聚类算法(TSS-FPCM)。

假设历史场景中有 M 个已知标签样本,将数据提取放在目标数据的后面,构成新的目标数据集 $X'=\{\boldsymbol{x}_k\mid k=1,2,\cdots,N,N+1,\cdots,N+M\}$,$\boldsymbol{x}_k\in R^d$,其中后 M 个数据为历史场景中的已知样本,根据数据集提出新的目标函数为

$$J=\sum_{i=1}^{C}\sum_{k=1}^{N}(\alpha u_{ik}^2+\beta t_{ik}^2)\,d_{ik}^2+\omega\left[\sum_{i=1}^{C}\sum_{k=N+1}^{N+M}(\alpha u_{ik}^2+\beta t_{ik}^2)\,d_{ik}^2+\sum_{i=1}^{C}\sum_{k=N+1}^{N+M}(u_{ik}-f_{ik})^2 d_{ik}^2\right]$$

s.t. $\alpha\geqslant 0,\ \beta\geqslant 0,\ \omega>0,\ 0\leqslant u_{ik},\ t_{ik}\leqslant 1$

$$\sum_{i=1}^{C}u_{ik}=1\ \forall k,\ \sum_{k=1}^{N+M}t_{ik}=1\ \forall i \tag{17}$$

不直接使用式(13)的目标函数而改用式(17)的目标函数,当参数 ω 趋于 0 的时候,前者相当于将 M 个源数据当作未知标签加入到目标领域中进行无监督混合 C 均值聚类,而后者则等于认为这些数据没有用处而舍弃。可以发现前者无法控制加入源数据后所可能造成的负迁移现象影响聚类结果,而后者则可以有效避免该情况。

最小化目标函数可以得到:

$$t_{ik}=\begin{cases}\left(\sum_{j=1}^{N}\frac{d_{ik}^2}{d_{ij}^2}+\sum_{j=N+1}^{N+M}\frac{d_{ik}^2}{\omega d_{ij}^2}\right)^{-1}, & k\leqslant N\\ \left(\sum_{j=1}^{N}\frac{\omega d_{ik}^2}{d_{ij}^2}+\sum_{j=N+1}^{N+M}\frac{d_{ik}^2}{d_{ij}^2}\right)^{-1}, & N<k\leqslant N+M\end{cases} \tag{18}$$

$$u_{ik}=\begin{cases}\left(\sum_{j=1}^{C}\frac{d_{ik}^2}{d_{jk}^2}\right)^{-1}, & k\leqslant N\\ \frac{1-\frac{1}{1+\alpha}\sum_{j=1}^{C}f_{jk}}{\sum_{j=1}^{C}\frac{d_{ik}^2}{d_{jk}^2}}+\frac{1}{1+\alpha}f_{ik}, & N<k\leqslant N+M\end{cases} \tag{19}$$

$$v_i=\frac{\sum_{k=1}^{N}(\alpha u_{ik}^2+\beta t_{ik}^2)\,x_k+\omega\sum_{k=N+1}^{N+M}\{\alpha u_{ik}^2+\beta t_{ik}^2+(u_{ik}-f_{ik})^2\}\,x_k}{\sum_{k=1}^{N}(\alpha u_{ik}^2+\beta t_{ik}^2)+\omega\sum_{k=N+1}^{N+M}\{\alpha u_{ik}^2+\beta t_{ik}^2+(u_{ik}-f_{ik})^2\}\,x_k},\ \forall i \tag{20}$$

2.3 改进的半监督迁移算法

在历史场景中,除了少量的标签信息,还有大量的未标记数据,这些数据量远远大于已标记数据,同样可以从中获取需要的信息来帮助当前场景。直接将大量未标记数据加入当前场景中进行聚类大大增加了计算量。

在历史场景中,为了减少计算量,可以使用一个“代表点”来表示一个类,而不仅仅是文献[11]中的聚类中心;这个点既可以是聚类中心,也可以是数据中的真实样本点,将庞大的数据变为有限的几个点。

为了能够有效地利用“代表点”,给定代表点集合 $X\hat{X}=\{\hat{x}_i\mid i=1,2,\cdots,C\}$,$C$ 表示聚类个数,重新定义新的距离函数为

$$\overset{\gg}{d}_{ik}^2 = \| \boldsymbol{x}_k - \boldsymbol{v}_i \|^2 + \gamma_1 \| \boldsymbol{x}_k - \hat{\boldsymbol{x}}_i \|^2 + \gamma_2 \| \boldsymbol{v}_i - \hat{\boldsymbol{x}}_i \|^2 \tag{21}$$

式中 γ_1 和 γ_2 为权重因子,用于调节历史中心的重要程度,将代表点作为有效信息迁移到当前场景中来。新的目标函数如式(22):

$$\begin{aligned} J = & \sum_{i=1}^{C}\sum_{k=1}^{N} (\alpha u_{ik}^2 + \beta t_{ik}^2) \overset{\gg}{d}_{ik}^2 + \\ & \omega \Big\{ \sum_{i=1}^{C}\sum_{k=N+1}^{N+M} (\alpha u_{ik}^2 + \beta t_{ik}^2) \overset{\gg}{d}_{ik}^2 + \\ & \sum_{i=1}^{C}\sum_{k=N+1}^{N+M} (u_{ik} - f_{ik})^2 \overset{\gg}{d}_{ik}^2 \Big] \end{aligned} \tag{22}$$

式中:$\alpha \geqslant 0$, $\beta \geqslant 0$, $\omega > 0$, $0 \leqslant u_{ik}$, $t_{ik} \leqslant 1$,

$$\sum_{i=1}^{C} u_{ik} = 1, \forall k, \sum_{k=1}^{N+M} t_{ik} = 1, \forall i。$$

为了获得其迭代表达式,利用拉格朗日极值优化表达式,首先构造 Lagrange 表达式:

$$\begin{aligned} Q = & \sum_{i=1}^{C}\sum_{k=1}^{N} (\alpha u_{ik}^2 + \beta t_{ik}^2) \overset{\gg}{d}_{ik}^2 + \\ & \omega \Big[\sum_{i=1}^{C}\sum_{k=N+1}^{N+M} (\alpha u_{ik}^2 + \beta t_{ik}^2) \overset{\gg}{d}_{ik}^2 + \sum_{i=1}^{C}\sum_{k=N+1}^{N+M} (u_{ik} - f_{ik})^2 \overset{\gg}{d}_{ik}^2 \Big] + \\ & \sum_{k=1}^{N+M} \lambda_k \Big(1 - \sum_{i=1}^{C} u_{ik}\Big) + \sum_{i=1}^{C} \theta_i \Big(1 - \sum_{k=1}^{N+M} t_{ik}\Big) \end{aligned} \tag{23}$$

式中 λ_k 与 θ_i 为 Lagrange 乘子。

令 $\partial Q / \partial V_i = 0$,解得:

$$v_i = \frac{\sum_{k=1}^{N} (\alpha u_{ik}^2 + \beta t_{ik}^2) x_k + \gamma_2 \sum_{k=1}^{N} (\alpha u_{ik}^2 + \beta t_{ik}^2) \hat{x}_i + \omega \Big[\sum_{k=N+1}^{N+M} (\alpha u_{ik}^2 + \beta t_{ik}^2 + (u_{ik} - f_{ik})^2) x_k + \gamma_2 \sum_{k=N+1}^{N+M} (\alpha u_{ik}^2 + \beta t_{ik}^2 + (u_{ik} - f_{ik})^2) \hat{x}_i \Big]}{(1 + \gamma_2) \Big[\sum_{k=1}^{N} (\alpha u_{ik}^2 + \beta t_{ik}^2) + \omega \sum_{k=N+1}^{N+M} (\alpha u_{ik}^2 + \beta t_{ik}^2 + (u_{ik} - f_{ik})^2) \Big]} \tag{24}$$

令 $\partial Q / \partial \lambda_k = 0$,可以得到:

$$\sum_{i=1}^{C} u_{ik} = 1 \tag{25}$$

令 $\partial Q / \partial u_{ik} = 0$,对于 $0 < k \leqslant N$ 可以解得:

$$u_{ik} = \frac{\lambda}{2\alpha \overset{\gg}{d}_{ik}^2} \tag{26}$$

将式(26)代入式(25),解得:

$$\frac{\lambda}{2\alpha} = \left(\sum_{i=1}^{C} \frac{1}{\overset{\gg}{d}_{ik}^2} \right)^{-1} \tag{27}$$

再将 λ 代回式(26),得到:

$$u_{ik} = \left(\sum_{j=1}^{C} \frac{\overset{\gg}{d}_{ik}^2}{\overset{\gg}{d}_{jk}^2} \right)^{-1} \tag{28}$$

同理,对于 $N < k \leqslant N+M$,可以求出:

$$u_{ik} = \frac{1 - \dfrac{1}{1+\alpha} \sum_{j=1}^{C} f_{jk}}{\sum_{j=1}^{C} \dfrac{\overset{\gg}{d}_{ik}^2}{\overset{\gg}{d}_{jk}^2}} + \frac{1}{1+\alpha} f_{ik} \tag{29}$$

合并式(28)和(29)可以得到最终表达式:

$$u_{ik} = \begin{cases} \left(\sum_{j=1}^{C} \dfrac{\overset{\gg}{d}_{ik}^2}{\overset{\gg}{d}_{jk}^2} \right)^{-1}, & k \leqslant N \\ \dfrac{1 - \dfrac{1}{1+\alpha} \sum_{j=1}^{C} f_{jk}}{\sum_{j=1}^{C} \dfrac{\overset{\gg}{d}_{ik}^2}{\overset{\gg}{d}_{jk}^2}} + \dfrac{1}{1+\alpha} f_{ik}, & N < k \leqslant N+M \end{cases} \tag{30}$$

使用同样得方法,可以求得 t_{ik} 的迭代表达式:

$$t_{ik} = \begin{cases} \left(\sum_{j=1}^{N} \dfrac{\overset{\gg}{d}_{ik}^2}{\overset{\gg}{d}_{ij}^2} + \sum_{j=N+1}^{N+M} \dfrac{\overset{\gg}{d}_{ik}^2}{\omega \overset{\gg}{d}_{ij}^2} \right)^{-1}, & k \leqslant N \\ \left(\sum_{j=1}^{N} \dfrac{\omega \overset{\gg}{d}_{ik}^2}{\overset{\gg}{d}_{ij}^2} + \sum_{j=N+1}^{N+M} \dfrac{\overset{\gg}{d}_{ik}^2}{\overset{\gg}{d}_{ij}^2} \right)^{-1}, & N < k \leqslant N+M \end{cases} \tag{31}$$

2.4 改进的半监督迁移算法描述

根据上一节的公式,ITSS-FPCM 的表述如下:

算法 1　ITSS-FPCM 算法

输入　前 N 个数据样本为目标数据,后 M 个为已知标签的历史数据的数据样本 $X' = \{\boldsymbol{x}_k \mid k = 1, 2, \cdots, N, N+1, \cdots, N+M\}$,聚类个数 C,最大迭代次数 L,当前迭代次数 $l = 1$,源数据类代表点 $\hat{X}$,相关参数 α、β、ω、γ_1 和 γ_2,阈值 ε。

输出　聚类中心 $\boldsymbol{v}_i$,隶属度矩阵 $\boldsymbol{u}_{ik}$ 和概率矩阵 $\boldsymbol{t}_{ik}$。

1)初始化聚类中心 $\boldsymbol{v}_i$,根据已知标签构造矩阵 $\boldsymbol{F}$,初始化目标函数 $J^{(l)} = 0$。

2)根据表达式(30)更新 $\boldsymbol{v}_{ik}$。

3)根据表达式(31)更新 $\boldsymbol{v}_{ik}$。

4)根据表达式(24)更新 $\boldsymbol{v}_i$。

5)$l = l + 1$,计算新的目标函数 $J^{(l)}$,如果 $J^{(l)} - J^{(l-1)} < \varepsilon$,或者 $l > L$ 跳到第 6),否则,跳到 2)。

6)聚类中心 v_i,隶属度矩阵 $\boldsymbol{v}_{ik}$ 和概率矩阵 $\boldsymbol{v}_{ik}$。

3　实验结果

为了验证算法的有效性,实验使用了人工数据

集、UCI 真实数据集以及文本数据集进行相关的实验验证。

在进行聚类结果评价时,选取了相关的 4 种聚类评价指标:正确率 AC(Accuracy)[18]、归一化互信息 NMI(normalized mutual information)[11,18]、芮氏指标 RI(Rand Index)[11,19] 和 F-measure[19]。4 个指标的值域均在 0 到 1,值越大表示聚类质量越好。

实验中选取了 LSSMTC[18]、Co-Clustering[20]、FPCM、TSC[12]、T-GIFP-FCM[11] 算法进行对比实实验;评价结果将进行 10 次计算取平均值。

3.1 人工数据集

为了模拟源场景和当前目标场景,实验使用文献[11]的方法:首先利用高斯函数生成相关的数据集,随机生成类别数为 3,每类 250 个样本点,每个样本点为两微的源场景数据,如图 1 所示。

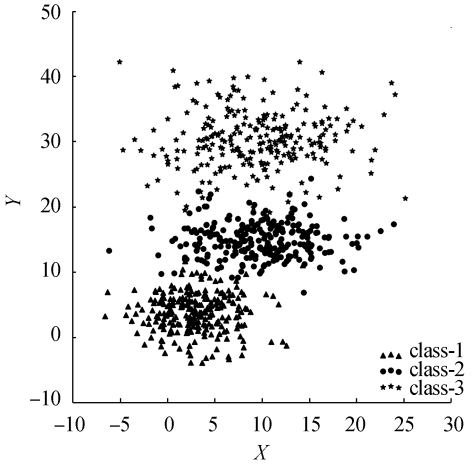
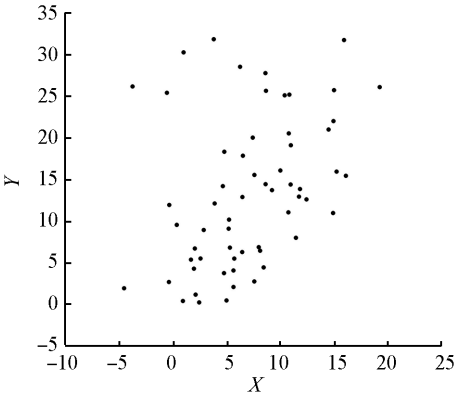


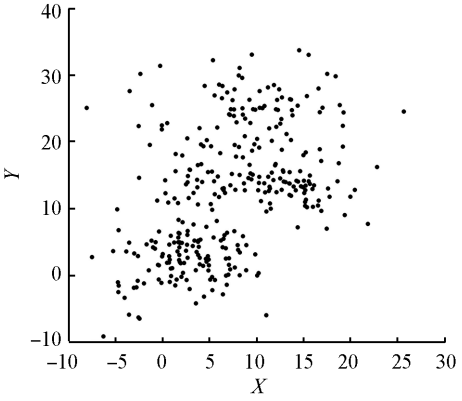
图 1 源数据

Fig.1 Source Dataset

如图 2 所示,同样利用高斯分布函数产生当前数据集 Set1 和 Set2 两个数据集;其中 Set1 每类样本数目为 20,如图 2(a)所示;Set2 每类样本数目为 100,再向其中加入高斯噪声构成,如图 2(b)所示。



(a)数据集 Set1



(b)数据集 Set2

图 2 目标数据集

Fig.2 Target dataset

两个数据集分别模拟当前的数据样本信息匮乏(数据不足)、充足(数据足够)但是受污染(有噪声)的不同情况下进行聚类。

实验时,SS-FPCM,TSS-FPCM,ITSS-FPCM 算法需要已知部分源标签,随机从源数据中抽取 3% 的样本作为已知标签数据进行实验,实验结果如表 1 所示,表格中“—”表示该数据集不满足算法运行的基本条件。

表 1 8 个算法在人工数据集的对比

Table1 Comparison of 8 algorithms on artificial data sets

数据集	评价指标	算法							
		LSSMTC	Co-Clustering	FPCM	TSC	T-GIFP-FCM	SS-FPCM	TSS-FPCM	ITSS-FPCM
Set1	F-measure	0.898 1	0.883 7	0.865 8	—	0.895 6	0.901 7	0.901 7	0.915 9
	RI	0.872 9	0.859 3	0.843 5	—	0.862 7	0.884 2	0.884 2	0.895 5
	AC	0.900 0	0.883 3	0.866 7	—	0.893 3	0.900 0	0.900 0	0.916 7
	NMI	0.706 7	0.743 4	0.656 1	—	0.736 4	0.732 2	0.732 2	0.769 8
Set2	F-measure	0.877 1	0.911 7	0.901 0	—	0.918 4	0.910 7	0.912 4	0.953 8
	RI	0.861 5	0.869 8	0.884 7	—	0.896 7	0.892 0	0.892 0	0.941 0
	AC	0.846 7	0.901 0	0.900 0	—	0.920 0	0.910 0	0.913 3	0.954 2
	NMI	0.718 7	0.770 5	0.761 6	—	0.801 6	0.781 0	0.788 0	0.844 4

从表 1 可以看出:

1)在 Set1 数据集中样本量很少,少量的源标签数据样本和其他信息都能够对目标数据产生正向的推动作用,从而达到较好的结果,SS-FPCM 与 TSS-FPCM 的结果验证了这一点;T-GITP-FCM 算法也可以得到很好的结果;

2)在有噪声的数据集 Set2 上,少量的标签不足以取得令人满意的效果,仍需要源数据的其他帮助,SS-FPCM 与 TSS-FPCM 算法的结果不如 T-GIFP-FCM 算法;说明 SS-FPCM 与 TSS-FPCM 算法在抗干扰方面存在不足;

3)改进后的 ITSS-FPCM 算法则在 Set1 和 Set2 上均取得了良好的聚类效果。说明当在数据信息不足,数据样本有限,数据受污染的时候,在有大量历史数据的帮助下迁移算法可以取得不错的效果,改进的 ITSS-FPCM 算法在抗噪声和干扰方面优于其他算法。

3.2　UCI 真实数据集

UCI 中的 Image Segment Data Set 是一个图片数据集,它由 7 个室外图像数据库中随机抽取,组成 7 个不同的类别,共 2 100 个样本数据,其中每个类别含有 300 个样本点。实验从数据中抽取 70%的数据作为源数据,剩下的构成目标数据进行实验,数据构成如表 2。

表 2　Image Segment 数据集构成情况

Table2　Composition of image segment data sets

数据类型	样本数	维数	类别
源数据	1 470	19	7
目标数据	630	19	7

算法在数据集的聚类结果如图 3 所示,从图中可以发现本文所提出的 ITSS-FPCM 算法在 4 个指标均取得了不错的结果,在准确率与 NMI 指标上有相对较大的优势,进一步验证了算法得有效性。

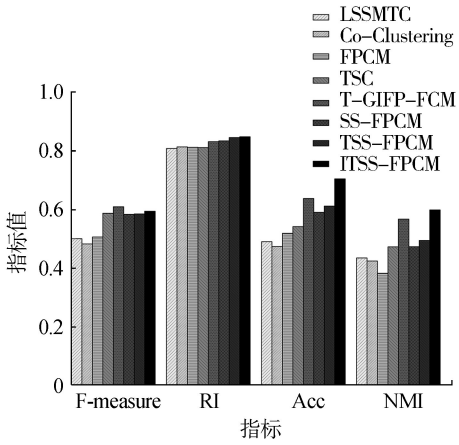


图 3　8 个算法在 Image Segment 数据集上的对比

Fig.3　Comparison of 8 algorithms on image segment data set

3.3　文本真实数据集

20NG(20Newsgroups)[12]是一个真实的新闻文本数据集,数据集收集了大约 2 万条新闻组,均匀地分布到 20 个不同的集合中,20 个小集合又可以分为 4 个大的类别,该数据集在大量迁移学习分类算法中被使用。

TDT2[21](NIST 话题检测与跟踪的语料库)共收集 1998 年上半年 6 个来源的数据,包含 2 个通讯社(APW,NYT),2 个电台节目(VOA,PRI)和 2 个电视节目(CNN,ABC),共 1 万多个样本数据。

Reuters-21578[21]语料库包含 21 578 个文件,放在 135 个文件夹下。

实验时分别对 3 个文本数据集抽取其中一部分类别,利用工具进行降维处理后构成新的数据集样本,数据具体构成如表 3 所示。

表 3　数据集构成情况

Table3　Composition of data sets

数据来源	数据类型	样本数	维数	类别
comp vs sci(20NG)	源数据	1 200	400	2
	目标数据	400	400	2
rec vs talk(20NG)	源数据	1 200	400	2
	目标数据	400	400	2
TDT2	源数据	1 800	400	6
	目标数据	600	400	6
Reuters-21578	源数据	800	400	4
	目标数据	400	400	4

聚类的结果如表 4 所示,结果中可以看到:

1) 利用迁移学习的 TSC、T-GIFP-FCM、TSS-FCM、ITSS-FCM 算法在效果上均优于非迁移学习型算法,表明迁移学习能够有效地提升聚类的性能;

2)仅对源数据少量标签数据直接使用的 SS-FPCM 算法和 TSS-FPCM 算法对当前场景的作用有限,不及能够利用更多信息的 TSC 迁移聚类和 T-GIFP-FCM 算法,但还是能够有效地提高聚类性能;

3) 本论文的 ITSS-FPCM 算法在大部分指标都优于其他算法,但是当源数据与目标数据相关性不大时,基于标签与代表点的直接迁移对当前场景帮助有限,不及 STC 算法的聚类效果,存在着局限性和适用范围的问题。

表 4 8 个算法在人工数据集的对比

Table4 Comparison of 8 algorithms on artificial data sets

数据集	评价指标	算法							
		LSSMTC	Co-Clustering	FPCM	TSC	T-GIFP-FCM	SS-FPCM	TSS-FPCM	ITSS-FPCM
sciSet1	F-measure	0.683 4	0.664 8	0.633 1	**0.768 8**	0.695 6	0.698 4	0.718 7	0.733 6
	RI	0.558 5	0.555 0	0.524 1	**0.645 0**	0.577 0	0.575 0	0.595 8	0.609 5
	AC	0.816 5	0.667 5	0.750 0	**0.770 0**	0.697 5	0.695 0	0.720 0	0.735 0
	NMI	0.134 1	0.102 1	0.118 9	**0.292 3**	0.148 3	0.109 8	0.134 2	0.156 4
rec vs talk	F-measure	0.686 7	0.639 4	0.698 0	0.882 7	0.890 7	0.831 1	0.846 9	**0.915 8**
	RI	0.580 3	0.539 5	0.576 9	0.792 1	0.803 7	0.720 4	0.740 9	**0.844 0**
	AC	0.705 3	0.642 5	0.697 5	0.882 5	0.890 0	0.832 5	0.847 5	**0.915 0**
	NMI	0.176 9	0.087 1	0.0993	0.463 7	0.487 3	0.349 2	0.375 0	**0.574 8**
TDT2	F-measure	0.6427	0.613 9	0.478 7	0.855 4	**0.8897**	0.821 4	0.825 3	0.885 8
	RI	0.782 8	0.747 3	0.682 5	0.907 0	0.9299	0.884 5	0.888 4	**0.930 0**
	AC	0.698 3	0.713 3	0.608 3	0.863 3	**0.8967**	0.833 3	0.835 0	0.888 3
	NMI	0.542 6	0.575 0	0.398 0	0.753 5	0.8093	0.719 9	0.721 7	**0.829 8**
Reuters-21578	F-measure	0.710 1	0.684 0	0.6361	0.824 7	0.8533	0.812 1	0.817 8	**0.860 8**
	RI	**0.812 5**	**0.715 3**	**0.6620**	**0.841 9**	**0.8658**	**0.832 3**	**0.837 6**	**0.870 9**
	AC	0.820 0	0.727 5	0.719 1	0.830 0	0.8550	0.815 0	0.820 0	**0.865 0**
	NMI	0.566 2	0.505 2	0.448 5	0.659 0	0.6430	0.616 2	0.624 2	**0.707 6**

4 结束语

本文将半监督学习思想应用到 FPCM 算法上，提出半监督 SS-FPCM 算法；迁移学习方面对算法进行非负迁移改进，得到 TSS-FPCM 算法，再利用“代表点”代替原始数据提出了改进的半监督的迁移聚类算法 ITSS-FPCM。在多种数据集上的实验验证表明，ITSS-FPCM 算法在性能上要好于 SS-FPCM 算法与 TSS-FPCM 算法。在数据量不足、数据被污染的情况下，ITSS-FPCM 算法能够提升聚类的性能；算法在源数据与目标数据相关不大时效果一般，下一步研究将会提取其他相关信息改善聚类性能，同时考虑参数的优化问题。

参考文献：

[1] 庄福振，罗平，何清，等. 迁移学习研究进展[J]. 软件学报，2015，26(1)：26-39.
ZHUANG Fuzhen，LUO Ping，HE Qing，et al. Survey on transfer learning research[J]. Journal of software，2015，26(1)：26-39.

[2] WEI Fengmei，ZHANG Jianpei，CHU Yan，et al. FSFP：transfer learning from long texts to the short[J]. Applied mathematics and information sciences，2014，8(4)：2033-2040.

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

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

[5] SAMANTA S，SELVAN A T，DAS S. Cross-domain clustering performed by transfer of knowledge across domains[C]//Proceedings of the 4th National Conference on Pattern Recognition，Image Processing and Graphics (NCVPRIPG). Jodhpur，India，2013：1-4.

[6] DAI Wenyuan，XUE Guirong，YANG Qiang，et al. Transferring naive Bayes classifiers for text classification[C]//Proceedings of the 22nd National Conference on Artificial Intelligence. Vancourver，British Columbia，Canada，2007，1：540-545.

[7] LIAO Xuejun，XUE Ya，CARIN L. Logistic regression with an auxiliary data source[C]//Proceedings of the 22nd International Conference on Machine Learning. New York，

NY, USA, 2005: 505-512.

[8] DAI Wenyuan, YANG Qiang, XUE Guirong, et al. Boosting for transfer learning[C]//Proceedings of the 24th International Conference on Machine Learning. Corvallis, Oregon, USA, 2007: 193-200.

[9] LUO Ping, ZHUANG Fuzhen, XIONG Hui, et al. Transfer learning from multiple source domains via consensus regularization[C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, California, USA, 2008: 103-112.

[10] DUAN Lixin, TSANG I W, XU Dong, et al. Domain adaptation from multiple sources via auxiliary classifiers[C]// Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada,, 2009: 289-296.

[11] 蒋亦樟, 邓赵红, 王骏, 等. 基于知识利用的迁移学习一般化增强模糊划分聚类算法[J]. 模式识别与人工智能, 2013, 26(10): 975-984.
JIANG Yizhang, DENG Zhaohong, WANG Jun, et al. Transfer generalized fuzzy c-means clustering algorithm with improved fuzzy partitions by leveraging knowledge[J]. Pattern recognition and artificial intelligence, 2013, 26(10): 975-984.

[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: lecture notes in computer science. Berlin Heidelberg: Springer, 2012, 7524: 789-803.

[13] 李昆仑, 曹铮, 曹丽苹, 等. 半监督聚类的若干新进展[J]. 模式识别与人工智能, 2009, 22(5): 735-742. LI Kunlun, CAO Zheng, CAO Liping, et al. Some developments on semi-supervised clustering[J]. Pattern recognition and artificial intelligence, 2009, 22(5): 735-742.

[14] PAL N R, PAL K, BEZDEK J C. A mixed c-means clustering model[C]//Proceedings of the 6th IEEE International Conference on Fuzzy Systems. Barcelona, Spain, 1997, 1: 11-21.

[15] BEZDEK J C, EHRLICH R, FULL W. FCM: The fuzzy c-means clustering algorithm [J]. Computers and geosciences, 1984, 10(2-3): 191-203.

[16] KRISHNAPURAM R, KELLER J M. The possibilistic C-means algorithm: insights and recommendations[J]. IEEE transactions on fuzzy systems, 1996, 4(3): 385-393.

[17] PEDRYCZ W. Algorithms of fuzzy clustering with partial supervision[J]. Pattern recognition letters, 1985, 3(1): 13-20.

[18] GU Quanquan, ZHOU Jie. Learning the shared subspace for multi-task clustering and transductive transfer classification[C]//Proceedings of the 2009 9th IEEE international conference on data mining. Miami, Florida, USA, 2009: 159-168.

[19] 杨燕, 靳蕃, KAME M. 聚类有效性评价综述[J]. 计算机应用研究, 2008, 25(6): 1630-1632, 1638.
YANG Yan, JIN Fan, KAME M. Survey of clustering validity evaluation [J]. Application research of computers, 2008, 25(6): 1630-1632, 1638.

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

[21] CAI Deng, HE Xiaofei, HAN Jiawei. Locally consistent concept factorization for document clustering[J]. IEEE transactions on knowledge and data engineering, 2011, 23(6): 902-913.

作者简介:

王跃,男,1990 年生,硕士研究生,主要研究方向为数据挖掘、计算智能。

杨燕,女,1964 年生,教授,博士生导师,主要研究方向为计算智能、数据挖掘、集成学习。主持国家自然科学基金项目 3 项,国家科技支撑计划项目 1 项,发表学术论文 130 余篇。

王红军,男,1977 年生,副研究员,主要研究方向为机器学习、深度学习、半监督学习。主持完成国家自然科学青年基金项目 1 项,主持国家自然科学基金项目 2 项,发表学术论文 30 余篇。