具备跨源域学习能力的模糊聚类算法研究*

吴文鹏, 刘　渊, 徐雁飞
(江南大学 数字媒体学院, 江苏 无锡 214122)

摘　要: 针对传统迁移学习聚类算法因单一源域到单一目标域且两者类别数必须一致的约束而达不到良好的聚类效果问题,提出了一种跨源域学习的聚类算法。该算法具有三大优点:a)仅扩大源域数目且取消了源域类别数的限定,算法可以自适应选择源域进行学习,所以算法的迁移学习能够得到较大的提升;b)由于所利用的源域知识不会暴露原数据,所以算法具有良好的源域数据隐私保护性;c)通过调节平衡参数可以使算法退化为传统的聚类算法,因此该算法的聚类性能是有所保障的。通过在模拟数据集和真实数据集上的实验,验证了该算法较之现有迁移学习聚类算法具有更好的迁移能力,且聚类性能及鲁棒性也有较大的提升。

关键词: 迁移学习; 跨源域学习; 隐私保护; 鲁棒性
中图分类号: TP391.41　　**文献标志码**: A　　**文章编号**: 1001-3695(2017)02-0357-04
doi:10.3969/j.issn.1001-3695.2017.02.008

Study of fuzzy clustering algorithm based learning ability across source domain

Wu Wenpeng, Liu Yuan, Xu Yanfei
(*School of Digital Media*, *Jiangnan University*, *Wuxi Jiangsu* 214122, *China*)

Abstract: As single source domain to a single target domain and the category number must be consistent constrained the traditional migration learning clustering algorithms, effect of clustering was not good, this paper proposed a clustering algorithm of learning across the source domain. This algorithm had three advantages. a) The new algorithm not only expands the number of the source domain but also canceled the limits of category number of the source domain, it could adaptively select the source domain to study, so the migration learning ability of the algorithm can get larger advance; b) As the source domain knowledge used by the algorithm would not expose the original data, so the algorithm had good data privacy protection of source domain. c) Adjusting the balance parameters could make the algorithm degrade into the traditional clustering algorithm, so clustering performance of the algorithm was guaranteed. Through the experiments on simulated data sets and real data sets, this paper verifies that the proposed algorithm has better migration ability compared with the traditional migration learning clustering algorithm, and the clustering performance and robustness have larger advance.

Key words: migration learning; learning across the source domain; privacy protection; robustness

0　引言

聚类是将事物按照某种相似性准则进行区分和分类的过程,在此过程中没有教师指导,因而聚类又称为无监督分类。最初的聚类方法属于硬划分范畴,如 K-means[1]。随着 Zadeh 的模糊集理论[2]的提出,硬聚类被推广发展为模糊聚类,基于划分的模糊聚类算法的研究也由此开始。典型的代表算法有模糊 C-均值算法(fuzzy C-means, FCM)[3~6]和极大熵聚类算法(maximum entropy clustering, MEC)[7~10]等。传统的基于划分模糊聚类算法在处理不同特征的数据样本时均有各自的性能,但此类方法的聚类策略具有一个共同的前提条件,即目标数据集必须是大量且无污染的。但在实际生活中由于各种因素的约束(如人力、企业数据保护),所采集的数据很难保证大量且无污染,这些问题都严重制约着传统模糊划分聚类算法的实用性。

为了解决上述问题,研究人员把迁移学习[11,12]的思想引入到聚类分析中,使传统的聚类算法具有学习能力[13,14],如 2009 年 Jiang 等在文献[13]中提出了一种半监督迁移成分分析(transfer component analysis, TCA)方法,该算法也被广泛应用到分类领域中,如 2015 年 Qian 等人[14]将迁移学习思想融入到极大熵聚类算法中(cross-domain, soft-partition clustering with diversity measure and knowledge reference, KT-CM),利用历史知识来指导当前数据集的聚类从而提高聚类算法的有效性和鲁棒性。上述研究人员提出的学习利用历史知识确实能够在一定程度解决传统聚类算法在处理数据量稀少及失真时的聚类性能,但他们也有一个共同的前提,即单一源域且源域与目标域的类别数必须是相同的,否则算法无法正常工作。但在实际生活中数据是相互关联且不断变化的,如小蝌蚪进化成青蛙的过程中特征就发生了巨大的变化。因此在面对上述类型的真实数据集时,现有的迁移学习聚类算法都会失效,所以现有单一迁移学习策略极大地阻碍了算法的实际应用。为了进一步提高迁移算法的学习能力,必须解决两类问题:a)源域数据集单一性的问题,这样可以将源域数据集扩展到多个,从而扩大源域数据集的范围;b)源域数据集与目标域数据集的类别数一致性的问题,这样可以使算法适应不断变化的真实数据集。

正对以上挑战,本文提出了多源域学习策略,即源域数据可以是多个,且源域与目标域的类别数可以是不同的。本文提

收稿日期: 2016-01-07; **修回日期**: 2016-06-29　**基金项目**: 江苏省自然科学基金资助项目(BK2011003)
作者简介:吴文鹏(1991-),男,硕士研究生,主要研究方向为社交网络、智能算法(wuwen_peng@163.com);刘渊(1967-),男,教授,主要研究方向为社交网络、网络安全;徐雁飞(1990-),女,硕士研究生,主要研究方向为社交网络.

出的多源域学习策略为源域类中心设置跨源域学习度量矩阵，这个矩阵是每个源域类中心对目标域的影响指数，通过不断迭代此迁移学习矩阵可以自适应地选择最合适的源域知识进行学习。本文将多源域学习策略融入到 FCM 算法中，得到具有跨多源域学习能力的聚类算法，称之为具备跨源域学习能力的模糊聚类算法(study of fuzzy clustering algorithm based the learning ability across the source domain, SFCA-LAD)。该算法具有以下三大特色：

a)本文算法取消了单一源域与源域类别数必须与目标域一致的限制，从而大大减小了源域数据采集的难度，也扩大了源域的范围，因此算法能够得到充分的学习。

b)本文算法通过不断迭代更新迁移学习率矩阵，可以使算法自适应地选择最具合适的源域类中心进行迁移学习，因此其迁移学习的能力得到了保证。并且通过调节平衡参数可以使算法退化为传统的 FCM 算法，因此本文算法的性能较之 FCM 算法只会得到提升而绝不差于传统 FCM 算法，有效地避免了负迁移的影响。

c)本文算法所利用的源域知识均是通过分析得出的高级知识(类中心)，不涉及原始数据，具有良好的源域隐私保护性。

1 相关知识

1.1 模糊 C-均值算法

在传统的聚类算法中有一类是通过不断迭代更新类中心来驱动空间的划分，其中 FCM 最为经典。FCM 的目标函数一般表示为

$$J_{\mathrm{FCM}}=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}^{m}\|x_j-v_i\|^2 \text{ s.t. } \mu_{ij}\in[0,1],\sum_{i=1}^{C}\mu_{ij}=1 \tag{1}$$

根据拉格朗日条件极值的相关求解策略给出了式(1)取得最优解时中心点 V 以及隶属度 U 的迭代公式如下：

$$v_i=\frac{\sum_{j=1}^{N}\mu_{ij}^{m}x_j}{\sum_{j=1}^{N}\mu_{ij}^{m}} \tag{2}$$

$$\mu_{ij}=\left[\sum_{k=1}^{C}\left(\frac{\|x_j-v_i\|^2}{\|x_j-v_k\|^2}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{3}$$

根据式(2)(3)两个迭代公式，迭代终止后获取隶属度矩阵 U，在对其进行去模糊化后就会得到最终的空间划分情况，根据隶属度矩阵可最终获取每个样本所对应的类标签。

1.2 迁移学习

1.2.1 迁移学习思想

迁移学习(transfer learning, TL)[11]是一种可以从现有的源域数据中抽取相关有益知识(信息)，用来帮助将来的新学习框架。迁移学习的目标是把从一个环境中学到的知识用来帮助新的目标域中的学习任务，以期增强最终的学习效果。目标域与源域具有一定相关性但又存在明显差异。

1.2.2 传统的迁移学习聚类算法

传统的迁移学习聚类算法将迁移学习的思想应用到聚类算法中，以解决目标域数据集稀少及失真时聚类算法获取不到令人满意的效果的问题。传统的迁移学习聚类算法的框架如图 1 所示，从图 1 中可以看出，传统的迁移学习聚类算法具有如下两大缺陷：a)只考虑了从单一源域到单一目标域之间的迁移过程，但实际生活中的数据是相互关联的，即多个源域与目标域相互关联，因此在面对真实数据时，传统的迁移学习算法就会失效；b)传统的迁移学习算法要求源域和目标域的类别数目必须一致，这就限制了迁移聚类算法的应用范围，也不能满足实际生活中日益变化的数据挖掘的要求。

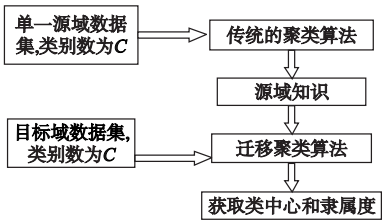

图1 迁移学习的学习框架

2 具备跨源域学习能力的模糊聚类算法

2.1 跨多源域学习策略

基于划分聚类算法的核心步骤就是迭代求出类中心和隶属度，其中类中心更为重要，类中心的好坏直接决定了聚类性能的好与坏。根据迁移学习的理论，源域数据集与目标域数据集之间存在一定的相似性，因此源域数据的类中心对于目标域类中心而言是一种很好的学习知识。但传统的迁移学习聚类算法只考虑均衡的单一源域的借鉴学习。如果能够自适应地迁移学习多个源域，这样就能更好地提高聚类算法迁移学习能力。基于此思想，首先提出跨多源域学习策略，设计约束项：

$$\sum_{k=1}^{C^s}\sum_{i=1}^{C^t}p_{ki}^{m}\|v_k^s-v_i^t\|^2 \tag{4}$$

其中：C^s 表示所有源域类别数的总和，C^t 表示目标域类中心数目；$v^s=(v^{s1}\cup v^{s2}\cup\cdots\cup v^{sn})$，$v_k^s\in v^s$ 表示 v^s 中第 k 个源域类中心，v_i^t 表示第 i 个目标域类中心；P 是 $C^s\times C^t$ 的矩阵，表示跨源域学习度量，$p_{ki}\in P$ 表示第 k 个源域类中心对第 i 个目标域类中心的影响程度，且 $p_{ki}\in[0,1]$，$\sum_{k=1}^{C^s}p_{ki}=1$。$p_{ki}=0$ 代表第 k 个源域类中心对第 i 个目标域类中心没有任何可借鉴学习的价值，$p_{ki}=1$ 表示第 i 个目标域类中心只借鉴学习第 k 个源域类中心。具体学习框架如图 2 所示。

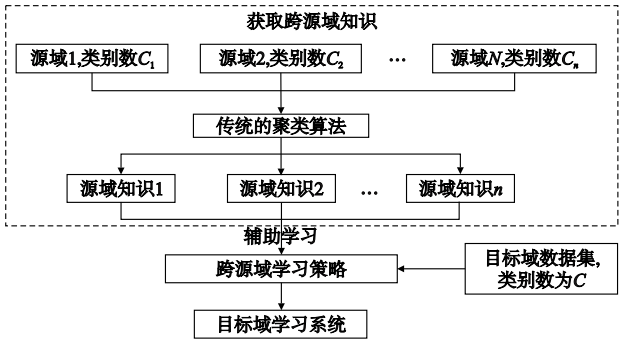

图2 跨源域学习策略框架

上述所提的跨多源域学习策略，式(4)可被应用于目前的基于划分的聚类算法中。跨域学习的核心参数是 P 矩阵，它是由源域类中心对目标域影响指数构成的，从矩阵的构成上可以看出，它跟隶属度矩阵的构成是相似的，隶属度矩阵是类中心对每个数据集的影响力度，聚类算法通过迭代来改变隶属度矩阵使算法收敛，同理算法也可以通过迭代调节 P 矩阵来自适应地选择最合适源域知识进行迁移学习，从而提高算法聚类效果。

2.2 具备跨源域学习能力的模糊聚类算法

将 2.1 节中的跨多源域学习策略融入到模糊 C-均值算法

中可以得到具有多源域学习能力的聚类算法,称为具备跨源域学习能力的模糊聚类算法。具体的表达形式如下:

$$J_{\text{SFCA-LAD}} = \sum_{i=1}^{C}\sum_{j=1}^{n}\mu_{ij}^{m} \| x_j - v_i^t \|^2 + \lambda \sum_{k=1}^{C^s}\sum_{i=1}^{C^t} p_{ki}^{m} \| v_k^s - v_i^t \|^2$$

$$\text{s.t}\quad \mu_{ij} \in [0,1], 1 \leqslant i \leqslant C, 1 \leqslant j \leqslant N, \sum_{i=1}^{C^t}\mu_{ij} = 1, \sum_{k=1}^{C^{vs}} p_{ki} = 1 \tag{5}$$

根据 Lagrange 条件极值最优解的方法可解出式(5)类中心、隶属度和跨源域学习度量的迭代公式。Lagrange 表达式为

$$L = \sum_{i=1}^{C}\sum_{j=1}^{n}\mu_{ij}^{m} \| x_j - v_i^t \|^2 + \lambda \sum_{k=1}^{C^s}\sum_{i=1}^{C^t} p_{ki}^{m} \| v_k^s - v_i^t \|^2 + \sum_{j=1}^{N}\alpha_j (1 - \sum_{i=1}^{C^t}\mu_{ij}) + \sum_{i=1}^{C_k^s}\beta_i (1 - \sum_{k=1}^{C^s} p_{ki}) \tag{6}$$

1)类中心迭代公式

$$\frac{\partial L}{\partial v_i^t} = 2\sum_{j=1}^{N}\mu_{ij}^{m}(x_j - v_i^t) + 2\lambda \sum_{k=1}^{C^s} p_{ki}^{m}(v_k^s - v_i^t) = 0 \tag{7}$$

解式(7)可得

$$v_i^t = \frac{\sum_{j=1}^{N}\mu_{ij}^{m} x_j + \lambda \sum_{k=1}^{C^s} p_{ki}^{m} v_k^s}{\sum_{j=1}^{N}\mu_{ij}^{m} + \lambda \sum_{k=1}^{C^s} p_{ki}^{m}} \tag{8}$$

2)隶属度迭代公式

$$\frac{\partial L}{\partial \mu_{ij}} = m\mu_{ij}^{m-1} \| x_j - v_i^t \|^2 - \alpha_j = 0 \tag{9}$$

则 $\mu_{ij} = \left(\frac{\alpha_j}{m \| x_j - v_i^t \|^2} \right)^{\frac{1}{m-1}}$,因为 $\sum_{p=1}^{C^t}\mu_{ij} = 1$,解得

$$\mu_{ij} = \left[\sum_{p=1}^{C^t} \left(\frac{\| x_j - v_i^t \|^2}{\| x_j - v_p^t \|^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{10}$$

3)跨源域学习度量迭代公式

$$\frac{\partial L}{\partial p_{ki}} = \lambda m p_{ki}^{m-1} \| v_k^s - v_i^t \|^2 - \beta_i = 0 \tag{11}$$

则 $p_{ki} = \left(\frac{\beta_i}{\lambda m \| v_k^s - v_i^t \|^2} \right)^{\frac{1}{m-1}}$,因为 $\sum_{h=1}^{C^s} P_{ki} = 1$,解得

$$p_{ki} = \left[\sum_{h=1}^{C^s} \left(\frac{\| v_k^s - v_i^t \|^2}{\| v_h^s - v_i^t \|^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{12}$$

根据迭代式(8)(10)和(12),最终可求得当前域数据集的最优类中心 V、隶属度 U 和跨源域学习度量矩阵 P。

2.3　具备多源域学习能力的模糊聚类算法步骤

具备多源域学习能力的模糊聚类算法可总结为以下六步。

输入:当前域数据集 data,目标域类别数 C,最大迭代次数 maxiter,迭代结束值 ε,多源域知识 V^s 和平衡参数 λ。

输出:类中心 V,隶属度 U 和跨源域学习度量矩阵 P。

1)获取多源域知识

对多源域数据集分别用传统的 FCM 算法进行聚类分析,得出所有源域类中心 V^s(如果已有源域类中心,可跳过此步)。

2)跨多源域学习聚类算法

a)初始化隶属度矩阵 $U(0)$,$P(0)$,将迭代次数置为 1;

b)根据最新的隶属度、跨源域学习度量和式(8)计算出最新的类中心 $V(n+1)$;

c)根据最新的类中心、跨源域学习度量和式(10)更新隶属度 $P(n+1)$;

d)根据最新的类中心、隶属度和式(12)更新跨源域学习度量矩阵 $P(n+1)$;

e)若 $\| U(n+1) - U(n) \|_F < \varepsilon$ 或迭代次数 n 达到 maxiter 时迭代结束,否则 $n = n+1$,并返回 b);

f)类中心 V,隶属度 U 和跨源域学习度量矩阵 P。

3　实验设置与结果分析

通过模拟数据及大量真实数据来验证本文算法较之传统的 FCM 和迁移聚类算法在处理信息缺失或失真的聚类任务时具有更好迁移能力且聚类性能更优。为进一步说明本文所提的多源域迁移学习策略的优越性,按照聚类的发展过程从每个阶段中选择几种典型的相关算法进行比较评价。在经典聚类算法中选择 FCM,在多任务聚类算法中选择基于共享子空间的多任务聚类算法(learning the shared subspace for multitask clustering,LSSMTC)[15]与组合 K-均值算法(combine K-means,CombKM);在迁移聚类算法中选择自学习聚类(self-taught clustering,STC)[13]与基于知识利用的迁移学习模糊划分聚类算法(transfer generalized fuzzy C-means clustering algorithm with improved fuzzy partitions by leveraging knowledge, T-GIFP-FCM)[3]。在实验中有关参数设置采用网格搜索法进行遍历寻优。FCM、STC 和 T-GIFP-FCM 采用原文献推荐的参数区间设置。SFCA-LAD 算法中的 λ 参数的寻优区间为[0.1:0.05:0.9]∪[1:1:10]∪[20:10:100]。

在与相关算法进行性能比较时,本文采用两种评价指标进行分析讨论:归一化互信息(normalized mutual information, NMI)[16]和芮氏指标(rand index,RI)[13],这两种指标取值范围均为[0,1],数值越高表示聚类性能越好。实验数据中 NMI-mean\ NMI-std(RI-mean\ RI-std)分别表示在最优参数下运行 10 次后 NMI 的均值与方差(RI 的均值与方差)。

3.1　模拟数据实验及结果分析

为了模拟真实生活中的数据,在源域数据中模拟设计了三个源域,分别为(S_1,S_2,S_3),源域的类别分别为 4,3,2 类,目标域数据为三类且数据集是稀少且存在污染的。每个源域都与目标域数据具有一些相似性也存在着比较大的差别。因此在面对这样的多源域且类别数不一致的数据集时,除了本文所提算法别的迁移算法都已失效,但为了突出本文算法的优越性,本文将源域 2(S_2)作为其余算法的迁移学习知识,这样其余算法就可以正常工作。模拟数据集具体参数分别描述如下:

1)多源域数据集　源域 1 为四类,每类 250 个点,第一类 E_1 = [5 3],Σ_1 = [12 0;0 12];第二类 E_2 = [18 3],Σ_2 = [12 0;0 22];第三类 E_3 = [19 23],Σ_3 = [15 0;0 15];第四类 E_4 = [31 18],Σ_4 = [5 0;0 5],源域 1 数据集简称为 S_1。源域 2 为三类,每类 250 个点,第一类 E_1 = [8 25],Σ_1 = [30 0;0 20];第二类 E_2 = [25 20],Σ_2 = [5 0;0 50];第三类 E_3 = [35 10],Σ_3 = [10 0;0 15],源域 2 数据集简称为 S_2。源域 3 为两类,每类 250 个点,第一类 E_1 = [11 13],Σ_1 = [25 0;0 7];第二类 E_2 = [9 29],Σ_2 = [30 0;0 20],源域 3 数据集简称为 S_3。多源域数据集如图 3 所示。

2)目标域数据集　目标域 1 分为三类,每一类有 20 个点,第一类 E_1 = [3 4],Σ_1 = [10 0;0 10];第二类 E_2 = [11 13],Σ_2 = [25 0;0 7];第三类 E_3 = [9 29],Σ_3 = [30 0;0 20],目标域 1 数据集简称为 D_1。目标域 2 分为三类,每一类有 125 个点,第一类 E_1 = [3 4],Σ_1 = [10 0;0 10];第二类 E_2 = [11 13],Σ_2 = [25 0;0 7];第三类 E_3 = [9 29],Σ_3 = [30 0;0 20],并且在第三类上加均值为 0,方差为 2 的干扰信息,使其失真,目标域 2 数据集简称为 D_2,如图 4 所示。

根据上述的模拟数据,五种算法的实验结果如表 1 所示。分析实验数据可得如下结论:

a)在 D_1 数据集上可以发现传统 FCM 算法聚类效果不如 T-GIFP-FCM 和 CFPCA 算法。这是因为在数据量稀少时,传统 FCM 算法无法根据当前域数据获得准确的类中心,因而无法取得满意的聚类效果,但 T-GIFP-FCM 和 CFPCA 算法有效利

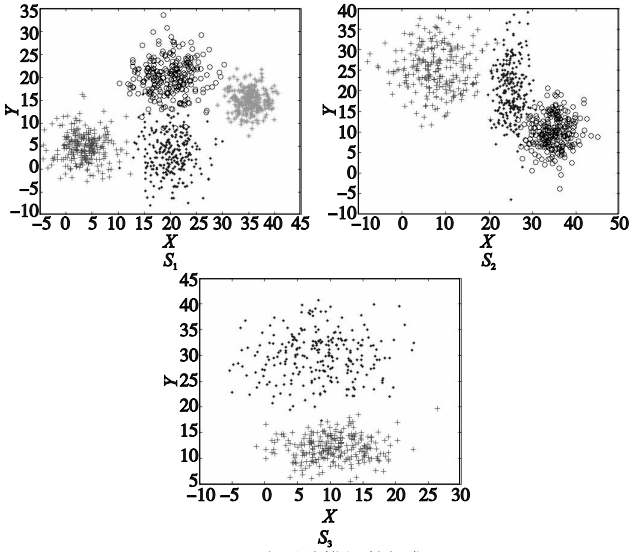

图3　多源域模拟数据集

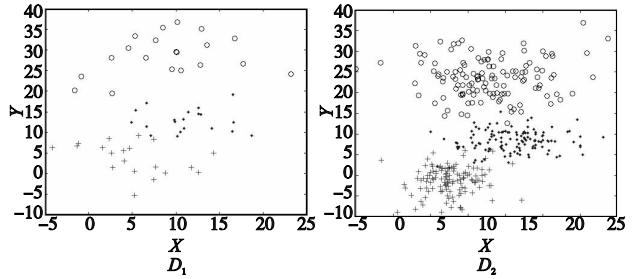

图4　目标域模拟数据集

用源域知识进行辅助学习,从而能够取得较好的聚类效果。进一步可以发现 T-GIFP-FCM 聚类性能不如 CFPCA 算法,因为 CFPCA 算法充分利用了三个源域知识,而 T-GIFP-FCM 算法因为其迁移学习策略所影响只能利用 S_2 源域知识,所以其迁移学习能力不如 CFPCA 算法。

b) D_2 数据集是相对充分但存在污染,从实验结果中可以发现 CFPCA 算法聚类性能最好,明显优于 T-GIFP-FCM 和 FCM 算法,而 T-GIFP-FCM 算法聚类性能相比于 FCM 算法略有提升,这是因为 D_2 数据集中的第三类数据严重失真,其类中心已不再明显,且源域 S_2 中不能提供此类数据的有效迁移学习的数据,因而 T-GIFP-FCM 算法的迁移能力得到了限制,而其聚类性能不如具备跨多源域学习能力的 CFPCA 算法。

c) 通过对实验结果的分析,STC 算法聚类性能好于 LSSMTC,而 LSSMTC 算法并不适合噪声数据的聚类任务,且 LSSMTC 和 STC 算法正常工作的前提是要得到源域数据集。

表 1　不同算法在模拟数据集上的性能对比

数据集	评价指标	算法				
		LSSMTC	FCM	STC	T-GIFP-FCM	SFCA-LAD
D_1	NMI-mean	0.789 9	0.819 7	0.815 1	0.848 1	0.907 5
	NMI-std	6.41E-17	1.73E-16	5.23E-16	0	0
	RI-mean	0.914 1	0.936 2	0.914 7	0.947 5	0.967 8
	RI-std	1.17E-16	1.73E-16	6.17E-16	0	0
D_2	NMI-mean	0.740 7	0.769 8	0.778 1	0.780 3	0.832 4
	NMI-std	1.17E-16	1.73E-16	6.36E-16	0	1.17E-16
	RI-mean	0.900 1	0.917 0	0.924 2	0.927 4	0.949 6
	RI-std	1.17E-16	0	1.36E-16	0	1.17E-16

3.2　真实数据集实验及结果分析

20 个新闻组(20Newsgroups,NG20)文本数据集收集大约 20 000 件新闻组文档,涵盖行政、体育等 20 个不同领域的新闻事件。NG20 可分为 4 个大类,每个大类包含多个子类。从以上特征看出 NG20 很好地反映了不同文本数据集所具有的特征。

本文从 NG20 中选择两类作为目标域数据集,选择两类、三类作为源域数据集, NG20 数据维数很高,有必要对它进行处理使其降维。为了满足目标域数据量严重不足的情况,从目标域子类中随机选取 250 个样本共 500 个样本作为目标域数据集,具体样本参数如表 2(a)(b)所示。具体的实验结果如表 3 所示。同样为了更好地突出本文算法,本文将源域 1 作为其余迁移算法的迁移学习知识。

表 2(a)　NG20 源域数据集说明

源数据集	数据组合	维度	样本容量	类别数
rec VS sci	rec. autos	350	750	2
	sci. med	350	750	
talk VS comp VS sci	talk. politics. misc	350	750	3
	comp. windows. x	350	750	
	sci. space	350	750	

表 2(b)　NG20 目标域数据集说明

目标数据集	数据组合	维度	样本容量	类别数
rec VS talk	rec. sport. baseball	350	250	2
	talk. politics. mideast	350	250	

表 3　不同算法算法 UCI 聚类性能对比

数据集	评价指标	算法				
		LSSMTC	FCM	STC	T-GIFP-FCM	SFCA-LAD
rec VS talk	NMI-mean	0.219 3	0.218 7	0.249 5	0.301 6	0.395 3
	NMI-std	0	0	0	0	1.84E-16
	RI-mean	0.608 6	0.601 4	0.614 4	0.608 3	0.674 1
	RI-std	0	1.17E-016	0	0	0

从表 3 的实验结果中可以看出,CFPCA 算法的聚类性能明显优于其他算法,这进一步说明本文算法可以很好地应用到真实数据集的聚类任务中。本文算法准确灵活地将通过总结多源域数据得到的知识应用到新的聚类任务中,从而提高聚类效果。这也再次验证了跨域利用源域知识能对当前域数据集的聚类任务起到很好的迁移学习指导作用。

4　结束语

本文针对现有的迁移学习聚类算法只能进行均衡单一源域学习,因而其迁移能力是有限的,为了提高算法的迁移能力,本文提出了跨多源域学习策略,并将其应用到模糊 C-均值算法中,从而得到具备跨源域学习能力的聚类算法。

然而,对于本文算法而言今后还有很多工作值得进一步研究和探讨,例如本文算法中初始使每个源域对目标域的影响程度都是一致的,今后可以考虑在每个源域前设置一个参数来表示对目标域的影响程度,这样设置更符合实际的生产应用。

参考文献:

[1] Alexander N G, Andrei Y Z. PCA and K-means decipher genome [C]//Principal Manifolds for Data Visualization and Dimension Reduction. Berlin:Springer-Verlag,2008: 309-323.

[2] Zadeh L A. Fuzzy sets[J]. Information and Control,1965,8(3): 338-353.

(下转第　　页)

PM2.5 浓度更加密切。经过权值优化后建立的模型与原数据建立的模型相比,前者在测试集上的泛化能力更好。同时如果对降维后的变量分别置 0 后,发现在训练集训练出的模型在测试集上的相对误差都有一定程度的增大。可见,采用 PSO-SVR(0-1)方法可以减少输入变量的维数,而采用 W-PSO-SVR 方法在降维的同时能保留对输入变量更多的信息,训练以后的结果过拟合程度较以前要低。

4 结束语

为了提高区域空气 PM2.5 浓度预测精度,减少对预测模型输入变量的复杂性和选择的盲目性,本文提出了一种在预测模型输入变量中添加[0,1]加权因子的支持向量回归机与粒子群优化算法相结合的浓度预测方法(W-PSO-SVR)。W-PSO-SVR 将支持向量回归机与粒子群优化算法结合,根据交叉验证的思想,通过不断搜索最佳加权因子,赋予输入变量不均等的权重,从而建立预测模型。预测效果表明,得到的预测精度与单纯的支持向量回归机模型和 0 或 1 加权因子的支持向量机模型相比有了明显的提高,并且 W-PSO-SVR 能较好地对模型的输入变量维数进行选择,减少了对模型输入变量选择的盲目性,从而起到了较好的降维效果。

参考文献:

[1] 张海涛,刘学飞,应宏,等.空气质量的混沌动力学分析[J].西南大学学报:自然科学版,2012,34(11):47-51.

[2] 张云海,孙财涛,杨洪斌.基于逐步回归分析方法的 PM10 浓度预报模型[J].环境科学与技术,2009,32(5):100-102.

[3] Russo A, Raischel F, Lind P G. Air quality prediction using optimal neural networks with stochastic variables[J]. Atmospheric Environment,2013,79:822-830.

[4] Yao Ling, Lu Ning, Jiang Sheng. Artificial neural network (ANN) for multi-source PM2.5 estimation using surface, MODIS, and meteorological data[C]//Proc of International Conference on Biomedical Engineering and Biotechnology. [S.l.]: IEEE Press, 2012: 1228-1231.

[5] Xie Hui, Ma Fei, Bai Qingyuan. Prediction of indoor air quality using artificial neural networks[C]//Proc of the 5th International Conference on Natural Computation. [S.l.]:IEEE Press,2009:414-418.

[6] Zheng Haiming, Shang Xiaoxiao. Study on prediction of atmospheric PM2.5 based on RBF neural network[C]//Proc of the 4th International Conference on Digital Manufacturing and Automation. [S.l.]: IEEE Press,2013:1287-1289..

[7] 李俭川,秦国川,温熙森,等.神经网络学习算法的过拟合问题及解决方法[J].振动、测试与诊断,2002,22(4):260-264.

[8] 邓乃杨,田英杰.数据挖掘中的新方法——支持向量机[M].北京:科学出版社,2006:224-235.

[9] 白鹏,张喜斌,张斌,等.支持向量机理论及工程应用实例[M].西安:西安电子科技大学出版社,2008:41-55.

[10] Sánchez A S, Nietob P J G, Fernándeza P R, *et al*. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain) [J]. Mathematical and Computer Modelling,2011,54(5-6):1453-1466.

[11] Xue Yansong, Wu Yang, Yu Le, *et al*. Support vector machine regression applied to MODIS data for PM 10 concentaration analysis [C]//Proc of the 2nd IITA International Conference on Geoscience and Remote Sensing. [S.l.]:IEEE Press,2010:51-54.

[12] Nieto P J G, Combarro E F, Del Díaz C J J, *et al*. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): a case study[J]. Applied Mathematics and Computation,2013,219(17):8923-8937.

[13] 秦侠,雷蕾,姚小丽.大气污染预测中提高 BP 网络泛化能力的方法[J].北京工业大学学报,2007,33(8):849-852.

[14] 赵宏,刘爱霞,王恺,等.基于 GA_ANN 改进的空气质量预测模型[J].环境科学研究,2009,22(11):1276-1281.

[15] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.

[16] 张利彪,周春光,马铭,等.基于粒子群算法求解多目标优化问题[J].计算机研究与发展,2004,41(7):1286-1291.

[17] Van Den Bergh F, Engelbrecht A P. A study of particle swarm optimization particle trajectories[J]. Information Sciences,2006,176(8):937-971.

(上接第 页)

[3] 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.

[4] Höppner F, Klawonn F, Kruse R. Fuzzy cluster analysis: methods for classification, data analysis, and image recognition[M]. New York: Wiley,1999.

[5] Miyamoto S, Mukaidono M. Fuzzy C-means as a regularization and maximum entropy approach [C]//Proc of the 7th International Fuzzy Systems Association World Congress. 1997:86-92.

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

[7] Karayiannis N B. Entropy-constrained learning vector quantization algorithms and their application in image compression [C]//Proc of SPIE, vol 3030. 1997: 2-13.

[8] Yu Jian, Shi Hongbo, Huang Houkuan, *et al*. Counter examples to convergence theorem of maximum-entropy clustering algorithm[J]. Science in China Series F: Information Sciences,2003,46(5): 321-326.

[9] Pal N R, Pal K. A mixed C-means clustering model [C]//Proc of the 6th IEEE International Conference on Fuzzy Systems. [S.l.]: IEEE Press,1997:11-21.

[10] Wang Shitong, Chung K F L, Deng Zhaohong, *et al*. Robust maximum entropy clustering algorithm with its labeling for outliers[J]. Soft Computing,2006,10(7): 555-563.

[11] Lisa T, Jude S, Trevor W, *et al*. Rule extraction for transfer learning [M]//Rule Extraction from Support Vector Machines. Berlin: Springer, 2008:67-82.

[12] Raina R, Battle A, Lee H, *et al*. Self-taughtlearning: transfer learning from unlabeled data[C]//Proc of the 24th International Conference on Machine Learning. New York: ACM Press,2007:759-766.

[13] Jiang Wenbo, Chung Fulai. Transfer spectral clustering[M]//Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2012:789-803.

[14] Qian Pengjiang, Sun Shouwei, Jiang Yizhang, *et al*. Cross-domain, soft-partition clustering with diversity measure and knowledge reference[J]. Pattern Recognition,2015,50:155-177.

[15] Dai Wenyuan, Yang Qiang, Xue Guirong, *et al*. Self-taught clustering [C]//Proc of the 25th International Conference on Machine Learning. New York: ACM Press,2008:200-207.