控制与决策
Control and Decision

增量学习的模糊风格K平面聚类

顾苏杭， 王士同

引用本文:
顾苏杭, 王士同. 增量学习的模糊风格K平面聚类[J]. 控制与决策, 2020, 35(9): 2081−2093.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0023

增量学习的模糊风格K平面聚类

顾苏杭[1,2†], 王士同[1]

(1. 江南大学 数字媒体学院，江苏 无锡 214122；
2. 常州轻工职业技术学院 信息工程与技术学院，江苏 常州 213164)

摘 要: 提出利用特征增量学习和数据风格信息双知识表达约束的模糊K平面聚类(ISF-KPC)算法. 为了获得更好的泛化性,聚类前利用高斯核函数对原输入特征进行增长式的特征扩维. 考虑数据集中来源于同一聚类的样本具有相同的风格,以矩阵的形式表达数据风格信息,并采用迭代的方式确定每个聚类的风格矩阵. 大量实验结果表明,双知识表达约束的ISF-KPC与对比算法相比能够取得竞争性的聚类性能,尤其在具有典型风格数据集上能够取得优异的聚类性能.

关键词: K平面聚类；风格信息；特征扩维；模糊聚类

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0023

引用格式: 顾苏杭, 王士同. 增量学习的模糊风格K平面聚类[J]. 控制与决策, 2020, 35(9): 2081-2093.

Incremental learning based fuzzy style K-plane clustering

GU Su-hang[1,2†], *WANG Shi-tong*[1]

(1. School of Digital Media，Jiangnan University，Wuxi 214122，China；2. School of Information Engineering and Technology，Changzhou Institute of Industry Technology，Changzhou 213164，China)

Abstract: A fuzzy K-plane clustering algorithm based on double knowledge representations about incremental feature learning and homogeneous style of data (ISF-KPC) is proposed. Before partitioning data samples into different groups, i.e., clusters, the feature augumentation is firstly conducted in an incremental manner based on the original inputs. Since data samples originating from a group share a same homogeneous style, the style information of each group, denoted by a style matrix, will be iteratively determined. By extensive experiments, the proposed ISF-KPC based on the double knowledge representations can obtain comparative clustering performance when compared to the adopted comparative clustering methods, especially it has its best clustering performance on the datasets with clearly homogeneous styles.

Keywords: K-plane clustering；style information；feature augumentation；fuzzy clustering

0 引 言

在许多实际应用中,来源于同一组或同一源的数据往往潜在或明显地表现出独特的数据风格[1-4]. 典型的应用包括癫痫脑电信号识别[5-6](采集于健康人群的脑电信号所表现出来的波形特征明显不同于患有癫痫的人群)、手写体识别[1,4,7-8](不同的作者写出来的字体风格完全不一样)、元音识别[1,8](英文中的元音发音互不相同). 这些数据所表现的风格信息完全有别于数据的物理特征,如距离、颜色和相似性. 另外,按照人的识别思维,人们习惯将具有相同风格的事物进行归类,因此考虑数据风格信息的识别模型符合数据的实际情况.

聚类是一种无监督数据分析,在未给出样本标签信息的情况下可将样本划分到不同的类别中. 聚类分析已广泛应用于识别、图像处理、计算机视觉以及文本分析[9-10]. 基于K平面的聚类算法以其新颖的聚类方式已获得越来越多科研人员关注[11-14],其主要思想是通过生成K个平面以代替基于中心点聚类算法中[15-17]的K个聚类中心点. 基于K平面的聚类算法与基于中心点的聚类算法相比,中心点浓缩了数据风格信息,平面更利于挖掘并表达数据风格信息. 另外,将数据风格信息用于数据分类并改进分类行为已取得相关研究进展[1-4]. Huang等[1]认为同一类数据拥有独特的数据风格,不同类数据之间的风格

收稿日期: 2019-01-04；**修回日期:** 2019-03-21.

基金项目: 国家自然科学基金项目(61572236, 61300151)；常州工业职业技术学院博士基金项目(BSJJ13101010)；常州工业职业技术学院新一代信息技术团队项目(YB201813101005).

责任编委: 刘宝碇.

†通讯作者. E-mail: gusuhang09@163.com.

相互区别,并以矩阵的形式计算每一类数据的风格信息用于改善分类模型的性能;Jiang等[2]提出的风格集中自动解码器能够提取鲁棒的图像风格特征代表,从而有效提高时尚、建筑以及漫画等图像的分类性能;Veeramachaneni等[3-4]提出一种风格约束的文本分类方法,通过可供选择的风格假设赋予文本中每个类(即每个风格)不同高斯密度,用以区别文本中不同类的风格特征. 该方法在文本数据集上的分类效果明显优于基于语义、词典(单一的数据物理特征)的文本分类方法. 然而,现有针对数据风格的研究成果主要集中于有监督学习算法,对将数据风格信息用于聚类分析并提高聚类性能的关注很少. 因此,针对以上所述内容,所提出算法主要集中于聚类分析且以K平面聚类算法为基础.

在聚类分析中,由于样本标签信息未知,与有监督数据分析分类算法相比,聚类算法获得相对较少的数据特征信息. 为了弥补可利用数据信息的不平衡,在聚类算法中对数据进行双知识表达,即对原输入特征增量式扩维进行增量学习并挖掘扩维后的数据风格信息. 双知识表达约束下的ISF-KPC分割数据样本时使得具有同一风格的样本成为一类. 本文主要贡献在于:1) 双知识表达中的增量学习体现了高斯型支持向量机(support vector machine, SVM) 中特征向高维映射的思想;2) 双知识表达中的数据风格信息有别于数据物理特征;3) 大量实验结果表明双知识表达能够很好地约束聚类行为,与所选对比算法相比,ISF-KPC至少能够取得竞争性的聚类性能,尤其在具有典型风格数据集上能够取得最好的聚类结果.

1 ISF-KPC

1.1 K平面聚类算法

由于ISF-KPC算法以K平面聚类算法为基础,本文首先简单介绍K平面聚类算法[11]. 考虑一个数据集$\boldsymbol{X}$包含N个样本,即$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_N]$且有$\boldsymbol{x}_i\in\boldsymbol{R}^d$. 根据$\boldsymbol{X}$,平面聚类算法($K$-plane clustering, KPC)首先生成K个平面,然后将每个样本划分到距离最近的平面中,即

$$\min_{\boldsymbol{\omega}_j,b_j}\frac{1}{2}\|\boldsymbol{X}_j\boldsymbol{\omega}_j+b_j\boldsymbol{e}\|;$$
$$\text{s.t. }\|\boldsymbol{\omega}_j\|=1. \tag{1}$$

其中:$\boldsymbol{X}_j$为与第j个聚类相对应的数据子集;$\boldsymbol{e}$为具有合适维度的列向量,其元素均为1;$\boldsymbol{\omega}_j$、b_j为聚类平面参数,可通过迭代收敛的方式求解所有聚类平面参数.

典型的K平面聚类为硬聚类算法,即一个样本只能属于一个K平面,而在ISF-KPC中引入模糊隶属度约束,即一个样本可被分割到多个聚类平面,因此式(1)可变为

$$\min_{\mu_{ij},\boldsymbol{\omega}_j,b_j}=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^m(\boldsymbol{X}_i\boldsymbol{\omega}_j+b_j)^2;$$
$$\text{s.t. }\sum_{j=1}^{k}\mu_{ij}=1,\ \|\boldsymbol{\omega}_j\|^2=1. \tag{2}$$

1.2 动机-双知识表达

ISF-KPC的聚类行为由双知识表达所约束,即增量学习和挖掘数据风格信息. 增量学习保证了ISF-KPC的泛化性,数据风格信息能够提高ISF-KPC的实际聚类精确度. 以下分别介绍这两个有关数据的知识表达.

与高斯型SVM相似,ISF-KPC通过增量学习将样本特征进行增长式扩维,在更高维空间中生成K个聚类平面并完成聚类分析. 首先,将原输入样本$\boldsymbol{x}_i$一一进行投射生成增强特征z_i^1,投射函数选择高斯函数,进而将所有增强特征组成增强节点$\boldsymbol{E}^1$,如图1所示. 将生成的增强节点嵌入到原特征空间中,此时原输入样本$\boldsymbol{X}$变为新输入样本$\boldsymbol{X}^1=[\boldsymbol{X},\boldsymbol{E}^1]$. 将新生成的输入样本同样利用高斯函数进行一一投射,并将生成的增强节点再次嵌入到原输入特征空间中,经过一系列特征扩维操作直至ISF-KPC取得最好聚类性能,此时原输入样本$\boldsymbol{X}$变为新输入样本$\boldsymbol{X}^s=[\boldsymbol{X},\boldsymbol{E}^s]$. 图1形象地解释了增量学习的过程,其中增强节点$\boldsymbol{E}^s$与原输入特征共同影响ISF-KPC的聚类性能. 大量实验结果表明,在特征增量学习过程中,s的取值范围为$1\sim3$,即特征扩维操作只需执行$1\sim3$次的情况下ISF-KPC能够取得最佳聚类性能. 与高斯型SVM将特征通过核化映射到无穷维相比,本文增量学习过程简单且易于执行,无需求解无穷维特征空间中的超平面参数.

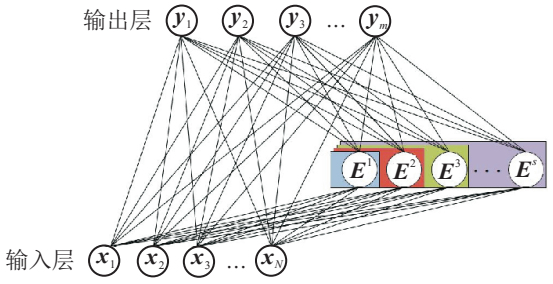


图1 增量学习

数据集中的每一类数据除了物理特征(如距离、颜色或相似性)外,还潜在或明显地拥有相应的风格,挖掘并利用数据风格信息可有效提高决策模型的性能[1-4]. 如图2为基于数据物理特征的决策模型(简称

为模型1)与基于数据风格信息的决策模型(简称为模型2)之间的区别. 其中,图2左边部分展现了包含若干字母的数据集,这些字母可组成不同的字体风格. 依据数据的相似性物理特征建立的模型1将具有相似外观的字母归为一类,如图2中间部分所示. 如果在建立决策模型的过程中考虑数据风格信息,模型2则将具有相同风格的字母归为一类,如图2右边部分所示. 模型2与模型1得到截然不同的归类结果,且模型2的归类结果一方面符合数据集中每一类数据具有各自的风格,另一方面也符合人们的认知习惯. ISF-KPC致力于挖掘并利用数据集中每一类数据风格信息改善其分类行为.

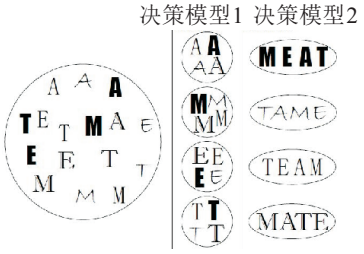


图 2　两种决策模型之间的区别

1.3　ISF-KPC目标函数及其优化

通过第2.2节动机描述可得双知识表达约束下的ISF-KPC目标函数为

$$J_{\text{ISF-KPC}}=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^{m}((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)^2+\lambda\sum_{j=1}^{k}\|\boldsymbol{A}_j-\boldsymbol{I}\|;$$

$$\text{s.t.}\ \sum_{j=1}^{k}\mu_{ij}=1,\ \|\boldsymbol{\omega}_j\|^2=1. \tag{3}$$

其中: $\boldsymbol{x}_i^s$为在原输入基础上连续执行增量学习后生成的新的输入; μ_{ij}为第i个样本归属第j个聚类的模糊隶属度,且有$\sum_{j=1}^{k}\mu_{ij}=1,\ \mu_{ij}>0,\forall i$; $\boldsymbol{A}_j\in R^{d\times d}$为与数据集中第$j$个聚类相对应的风格矩阵,通过风格矩阵$\boldsymbol{A}_j$迭代地挖掘每个聚类的数据风格信息; λ为风格调节参数,用于惩罚过度的数据风格信息,可人为确定或通过网格搜索结合交叉验证的方法确定其值,第2.1节会给出根据具体数据集确定其值的指导方法,另外,当$\lambda\to\infty$时,数据风格信息不再约束ISF-KPC聚类行为,当λ值设置过小时,ISF-KPC会过度地使用数据风格信息; $\boldsymbol{I}$为具有合适维度的单位矩阵. 大量的实验结果表明,当数据集中每一类数据具有典型数据风格时,风格矩阵$\boldsymbol{A}_j$中的元素值趋向于某一相对较大值,反之,风格矩阵$\boldsymbol{A}_j$中的元素值(除对角线元素)趋向于0.

与KPC算法相比,ISF-KPC具有以下显著区别: 1) ISF-KPC的目标函数中嵌入双知识表达,如式(3)所示,ISF-KPC的聚类行为受双知识表达约束,尤其数据风格信息$\boldsymbol{A}_j$能够有效地保证ISF-KPC的聚类精确度; 2) ISF-KPC目标函数带有正则项$\|\boldsymbol{A}_j-\boldsymbol{I}\|_F^2$($\|\cdot\|_F^2$为Frobenius范数),其作用于约束数据风格信息,由于该正则项基于数据风格信息,显著区别于SVM目标函数中正则项(能够很好地优化SVM模型性能).

1.4　ISF-KPC参数优化

通过ISF-KPC目标函数(式(3))可知,所提出聚类算法ISF-KPC共涉及5个参数,即样本模糊隶属度μ_{ij}、聚类平面参数$\boldsymbol{\omega}_j$和b_j、风格调节参数λ、风格矩阵$\boldsymbol{A}_j$. 其中: λ在第2.1节给出确定其值的指导方法, μ_{ij}、$\boldsymbol{\omega}_j$、b_j、$\boldsymbol{A}_j$四个参数通过以下两个独立程序迭代地确定其值.

1) 模糊K平面参数学习.

固定所有风格矩阵$\{\boldsymbol{A}_j\}(j=1,2,\ldots,k)$以及模糊隶属度矩阵,目标函数变为

$$J'_{\text{ISF-KPC}}=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^{m}((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)^2,$$

$$\text{s.t.}\ \|\boldsymbol{\omega}_j\|^2=1. \tag{4}$$

初始化$\{\boldsymbol{A}_j\}$使得任意$\boldsymbol{A}_j$为单位矩阵, $\boldsymbol{U}$可通过随机赋值或fuzzy C means (FCM)算法[16]进行初始化,且有$\sum_{j=1}^{k}\mu_{ij}=1,\mu_{ij}>0,\forall i$. 与式(1)相比,此时目标函数优化问题已转换成求解标准KPC中K个聚类平面参数问题[11]. 因此可通过求解特征值问题实现K平面参数学习. 考虑拉格朗日函数将式(4)转换为

$$L'(\boldsymbol{\omega}_j,b_j,\xi_j)=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)^2-\sum_{j=1}^{k}\xi_j(\boldsymbol{\omega}_j^{\mathrm{T}}\boldsymbol{\omega}_j-1). \tag{5}$$

将式(5)分别对$\boldsymbol{\omega}_j$、b_j的偏导求解,有

$$\frac{\partial L'}{\partial\boldsymbol{\omega}_j}=0,\ \frac{\partial L'}{\partial b_j}=0. \tag{6}$$

进而得到

$$\boldsymbol{D}_j\boldsymbol{\omega}_j=\xi_j\boldsymbol{\omega}_j, \tag{7}$$

$$b_j\sum_{i=1}^{k}\mu_{ij}^{m}=\boldsymbol{U}(:,j)^{\mathrm{T}}(\boldsymbol{X}^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j. \tag{8}$$

其中

$$\boldsymbol{D}_j=\frac{\left(\sum_{i=1}^{N}\mu_{ij}^{m}\boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{x}_i^s\right)\left(\sum_{i=1}^{N}\mu_{ij}^{m}(\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\right)}{\sum_{i=1}^{N}\mu_{ij}^{m}}-$$

$$\sum_{i=1}^{N}\mu_{ij}^{m}\boldsymbol{A}_j^{\mathrm{T}}\boldsymbol{x}_i^s(\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j, \tag{9}$$

$$\boldsymbol{D}_j=\frac{\boldsymbol{A}_j^{\mathrm{T}}(\boldsymbol{X}^s)^{\mathrm{T}}\boldsymbol{U}(:,j)\boldsymbol{U}(:,j)^{\mathrm{T}}\boldsymbol{X}^s\boldsymbol{A}_j}{\boldsymbol{e}^{\mathrm{T}}\boldsymbol{U}(:,j)}-$$
$$\boldsymbol{A}_j^{\mathrm{T}}(\boldsymbol{X}^s)^{\mathrm{T}}\Lambda\boldsymbol{X}^s\boldsymbol{A}_j. \tag{10}$$

$\boldsymbol{e}$为具有合适维度的列向量,其元素均为1;Λ为对角矩阵,其对角元素分别为$\mu_{1j}^m,\mu_{2j}^m,\dots,\mu_{Nj}^m$. 根据文献[11]求解KPC中$K$个聚类平面参数$\{\boldsymbol{\omega}_j\}$与$\{b_j\}$的方法,式(7)中$\xi_j$为$\boldsymbol{D}_j$的最小特征值,且$\boldsymbol{\omega}_j$为与最小特征值$\xi_j$相对应的特征向量,ISF-KPC中$K$个聚类平面参数$\{\boldsymbol{\omega}_j\}$即可求出,此时$b_j$为

$$b_j=\frac{\boldsymbol{U}(:,j)^{\mathrm{T}}(\boldsymbol{X}^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j}{\sum_{i=1}^{N}\mu_{ij}^m}. \tag{11}$$

至此,固定$\{\boldsymbol{A}_j\}$和$\boldsymbol{U}$,可求出K个聚类平面参数$\{\boldsymbol{\omega}_j\}$、$\{b_j\}$. 此时,在$\{\boldsymbol{A}_j\}$、$\{\boldsymbol{\omega}_j\}$、$\{b_j\}$都已知的情况下关于目标函数(如式(3)所示)的优化问题可转变为

$$J''_{\text{ISF-KPC}}=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^m((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)^2,$$
$$\text{s.t.}\ \sum_{j=1}^{k}\mu_{ij}=1. \tag{12}$$

同样考虑拉格朗日函数,式(12)变为

$$L''(\mu_{ij},\varepsilon_i)=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^m((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)^2-$$
$$\sum_{i=1}^{N}\varepsilon_i\Big(\sum_{j=1}^{k}\mu_{ij}-1\Big). \tag{13}$$

求解样本模糊隶属度可参照FCM算法[16],计算公式如下:

$$\mu_{ij}=\frac{1}{\sum_{j'=1}^{k}\Big(\frac{(\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j-b_j}{(\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_{j'}\boldsymbol{\omega}_{j'}-b_{j'}}\Big)^{\frac{2}{m-1}}}. \tag{14}$$

2) 数据风格信息学习. 当单独的模糊K平面参数学习程序结束后,即参数$\{\boldsymbol{\omega}_j\}$、$\{b_j\}$、$\boldsymbol{U}$已知,此时目标函数的优化问题转变为

$$J'''_{\text{ISF-KPC}}=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^m((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)^2+$$
$$\lambda\sum_{j=1}^{k}\|\boldsymbol{A}_j-\boldsymbol{I}\|_F^2. \tag{15}$$

由式(15)可知,$J'''_{\text{ISF-KPC}}$与k个独立的风格矩阵相关. 考虑拉格朗日函数,式(15)变为

$$L'''(\boldsymbol{A}_j)=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^m((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)^2+$$
$$\lambda\sum_{j=1}^{k}\|\boldsymbol{A}_j-\boldsymbol{I}\|_F^2. \tag{16}$$

很明显,式(16)是一个凸函数问题,通过求偏导的方法,有

$$\frac{\partial L'''}{\partial\boldsymbol{A}_j}=0, \tag{17}$$

很难求出所有风格的矩阵$\{\boldsymbol{A}_j\}$,参照文献[18],可通过迭代的方式求解$\{\boldsymbol{A}_j\}$. 首先定义中间变量υ_i并令其为

$$\upsilon_i=(\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j, \tag{18}$$

此时每个风格矩阵$\boldsymbol{A}_j$固定为单位矩阵. 式(16)变为

$$L'''(\boldsymbol{A}_j)=\sum_{i=1}^{N}\sum_{j=1}^{k}\mu_{ij}^m\upsilon_i((\boldsymbol{x}_i^s)^{\mathrm{T}}\boldsymbol{A}_j\boldsymbol{\omega}_j+b_j)+$$
$$\lambda\sum_{j=1}^{k}\|\boldsymbol{A}_j-\boldsymbol{I}\|_F^2. \tag{19}$$

由式(17)可得

$$\boldsymbol{A}_j=\boldsymbol{I}-\frac{1}{\lambda}\sum_{i=1}^{N}\mu_{ij}^m\upsilon_i\boldsymbol{x}_i^s\boldsymbol{\omega}_j. \tag{20}$$

在式(18)~(20)之间迭代计算$\boldsymbol{A}_j$与υ_i,直到满足$\sum_{i=1}^{N}\|\upsilon_i^{p+1}-\upsilon_i^p\|^2<\varphi$或达到最大迭代次数$P$[18]. 其中$\varphi$为一阈值,可根据实际实验结果人为设定.

根据模糊K平面参数学习和数据风格信息学习两个独立的程序可迭代计算出$\boldsymbol{U}$、$\{\boldsymbol{\omega}_j\}$、$\{b_j\}$、$\boldsymbol{A}_j$四个参数直到满足$\sum_{j=1}^{k}\|\boldsymbol{A}_j^{h+1}-\boldsymbol{A}_j^h\|^2<\theta$或达到最大迭代次数$H$,其中$\boldsymbol{A}_j^h$为第$h$次迭代过程中第$j$个聚类的风格矩阵.

1.5 ISF-KPC算法及其复杂度分析

通过第1.4节对ISF-KPC目标函数优化问题的描述,可得到以下算法过程描述.

输入: 给定数据集$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_N]^{\mathrm{T}}$,每个样本$\boldsymbol{x}_i\in R^d$,聚类数$k$,双知识表达中的增量学习执行次数$S$,所选高斯函数核宽度$\sigma$,风格调节参数$\lambda$,最大迭代次数$P$和$H$,迭代终止条件$\varphi$和$\theta$,模糊隶属度指数$m$;

输出: 各聚类平面参数$\{\boldsymbol{\omega}_j\}$与$\{b_j\}$,风格矩阵$\{\boldsymbol{A}_j\}$,模糊隶属度矩阵$\boldsymbol{U}$,预测标签集$\boldsymbol{Y}$.

step 1: 设定$s=1,\boldsymbol{X}^s=\boldsymbol{X}^0$.

step 2: 执行双知识表达中的增量学习,生成增强节点$\boldsymbol{Z}^s=[z_1^s,z_2^s,\dots,z_N^s]^{\mathrm{T}}$,继而生成新的输入$\boldsymbol{X}^s=[\boldsymbol{X}^{s-1},\boldsymbol{Z}^s]^{\mathrm{T}}$.

step 3: 设定$h=0$.

step 4: 循环直至程序收敛或达到最大迭代次数H.

step 4.1: 初始化$\{\boldsymbol{A}_j\}$、$\boldsymbol{U}$.

step 4.2: $h = h + 1$.

step 4.3: 利用式(7)和(11)求解$\{\boldsymbol{\omega}_j\}$、$\{b_j\}$.

step 4.4: 利用式(14)确定$\boldsymbol{U}$.

step 4.5: 迭代确定$\{\boldsymbol{A}_j\}$.

step 4.5.1: 设定$p = 1$;

step 4.5.2: 循环直至程序收敛或达到最大迭代次数P;

step 4.5.3: 利用式(18)和(20)确定$\{\boldsymbol{A}_j\}$.

根据ISF-KPC算法流程描述可得出以下算法复杂度分析(每个步骤均考虑最高阶情况下的复杂度): step 2中需要计算每一对样本间的距离,因此step 2占用的复杂度为$O(N^2)$. 对于step 4.1中初始化操作,如果考虑模糊隶属度矩阵$\boldsymbol{U}$中每个元素随机赋值且$\sum_{j=1}^{k}\mu_{ij} = 1, \mu_{ij} > 0, \forall i$,则step 4.1占用的复杂度为$O(2N)$. 对于step 4.3,在确定各个聚类平面参数$\boldsymbol{\omega}_j$、$b_j$前需要由式(10)计算$\boldsymbol{D}_j$,该操作相应的复杂度为$O(d^2N + 2dN^2 + d^3)$. $\boldsymbol{\omega}_j$为与$\boldsymbol{D}_j$最小特征值相对应的特征向量,因此确定$\boldsymbol{\omega}_j$占用的复杂度为$O(d^3)$. 根据式(11)可确定b_j,占用的复杂度为$O(d^3 + dN^2)$,因此当考虑最高阶时step 4.3占用的复杂度大小为$O[k(3d^3 + 3dN^2 + d^2N)]$. 根据式(14)可确定$\boldsymbol{U}$中每个元素,step 4.4占用的复杂度为$O(d^4N)$. step 4.5每一次迭代计算$\boldsymbol{A}_j$的过程中需要首先由式(18)确定$v_i$,该操作占用的复杂度为$O(d^3)$. 另外,根据式(20),需要$O(N^2d + 2d^2)$复杂度计算$\boldsymbol{A}_j$,因此step 4.5计算$\{\boldsymbol{A}_j\}$的复杂度为$O[kP(dN^2 + d^3)]$. 综上所述,ISF-KPC整个算法复杂度为$O[SH(d^4N + 3dN^2 + kPdN^2)]$. 在迭代次数有限的情况下,ISF-KPC算法复杂度主要与样本维度和样本总数相关. 因此ISF-KPC适合于样本维度和样本总数适当情况下的聚类分析.

2　实验与分析

本节将在人造数据集以及具有典型数据风格的真实数据集上验证所提ISF-KPC的聚类性能,并通过与其他算法的比较表明ISF-KPC中双知识表达的有效性.

2.1　对比算法及参数设置

由于ISF-KPC以KPC算法[11]为基础,将KPC及其模糊化版本F-KPC作为对比算法. 为了验证双知识表达的有效性,将ISF-KPC的简易版本ISF-KPC_0,即仅考虑单一知识表达(数据风格信息学习)作为对比算法之一. ISF-KPC中引入模糊隶属度约束,并将聚类中心点替换成聚类平面,选择典型的基于中心点的聚类算法K-medoids[15]和FCM[16]作为对比算法. 由于affinity propagation (AP)[19-20]和density-based spatial clustering of applications with noise (DBSCAN)[21-22]两种算法能够识别具有任意形状的数据,将这两种算法也作为对比算法.

由式(3)和增量学习过程可知,ISF-KPC共涉及6个参数,分别为聚类平面参数$\{\boldsymbol{\omega}_j\}$、$\{b_j\}$、模糊隶属度矩阵$\boldsymbol{U}$、风格矩阵$\{\boldsymbol{A}_j\}$、风格调节参数$\lambda$和增量学习高斯函数中的核宽度$\sigma$. $\{\boldsymbol{\omega}_j\}$、$\{b_j\}$、$\boldsymbol{U}$及$\{\boldsymbol{A}_j\}$可由模糊$K$平面参数学习与数据风格信息学习两个独立的程序根据真实数据集迭代确定,因此ISF-KPC主要确定λ和σ两个参数. 根据大量的实验结果可给出以下关于这两个参数的推荐设置: 参数σ的搜索范围为$\{10^{-2}, 10^{-1}, \ldots, 10^2, 10^3\}$,对于$\sigma$的每一个值,增量学习后所有新生成的数据间的平均欧氏距离为$\bar{D} = \dfrac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}$. 其中: N为样本总数, d_{ij}为第i个样本与第j个样本之间的欧氏距离. 由此可设置参数λ的值等于$\bar{D}$的量级,或在$\bar{D}$的量级附近进行搜索. 对于λ和σ,均采用网格搜索相结合交叉验证的方法进行确定[5,23].

关于其他对比算法,ISF-KPC_0并没有考虑增量学习,其涉及的参数$\{\boldsymbol{\omega}_j\}$、$\{b_j\}$、$\boldsymbol{U}$、$\{\boldsymbol{A}_j\}$以及$\lambda$均可参考ISF-KPC进行设置. KPC参数$k$设置为数据集中聚类数. F-KPC参数$k$设置为数据集中聚类数,模糊隶属度矩阵$\boldsymbol{U}$及模糊隶属度指数可参照FCM算法. K-medoids参数k设置为数据集中聚类数,FCM算法采用默认设置. AP算法的聚类性能主要受参数参考度PR影响,参考文献[19],在样本相似度中值附近采用网格搜索结合交叉验证的方法确定PR. DBSCAN算法主要受参数样本邻域阈值Eps和样本邻域内样本个数阈值MinPts影响,在文献[22, 24]推荐设置值附近采用网格搜索结合交叉验证的方法分别确定Eps和MinPts.

对于KPC、F-KPC以及K-medoids三种算法,分别运行30次后取平均结果,其他算法运行10次后取平均结果. 所有算法均在Matlab平台上运行,电脑配置为: 3.6 GHz且Intel(R) Core(TM) i7-4790 CPU, 8 G内存, 64位Windows 10操作系统.

2.2 数据集

本文在人造数据集上视觉地展示ISF-KPC算法的聚类性能,在真实数据集上验证ISF-KPC的实际聚类性能.

图3为本文采用的人造数据集,其中SD1、SD2、SD3中每一类数据对应典型的形状,意味着这3个数据集包含明显的数据风格. 由于SD4中有一类数据为高斯随机分布,SD5中4类数据对应同样的形状,SD6中3类数据都为高斯随机分布,因此这3个数据集并不包含明显的数据风格. 人造数据集详细配置信息如表1所示.

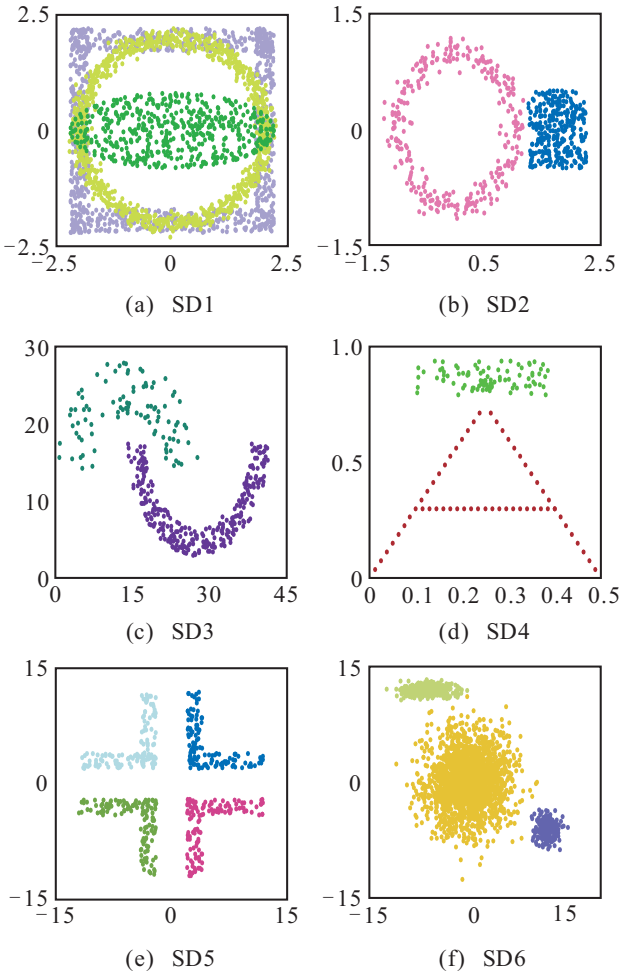


图3 人造数据集

表1 人造数据集详细配置

数据集	样本数	聚类数
SD1	2501	3
SD2	600	2
SD3	373	2
SD4	150	2
SD5	504	3
SD6	2600	3

本文在真实数据集上展示ISF-KPC独特的聚类性能,且每一个真实数据集具有明显的数据风格. 如引言所述,选取癫痫脑电信号识别、手写体识别以及元音识别作为3个案例.

对于癫痫脑电信号[5-6],其原始信号和经核化主成分分析(kernel principal component analysis, K-PCA)降维后的特征分别如图4和图5所示. 由图4和图5可见,正常人群的脑电信号(A组和B 组)明显区别于患癫痫人群的脑电信号(C组、D组和E组),即使属于同一人群,所表现出的脑电信号也相互区别,如A组和B组所示. 实验中分别将A组与B组,B组与D组,B组、D组与E组组成真实数据集,并分别命名为EEG-D1、EEG-D2和EEG-D3. 3个真实数据集的详细配置如表2所示,每个数据集都包含不同的数据风格,且每个聚类对应的数据风格各不相同.

对于手写体识别[1,4,7-8],实验中选取Chinese academy of science institute of automation (CASIA)官网公布的手写体数据集[7],其中部分数据如图6所示. 由此可见,出自每一位作者的手写体均相互区别,

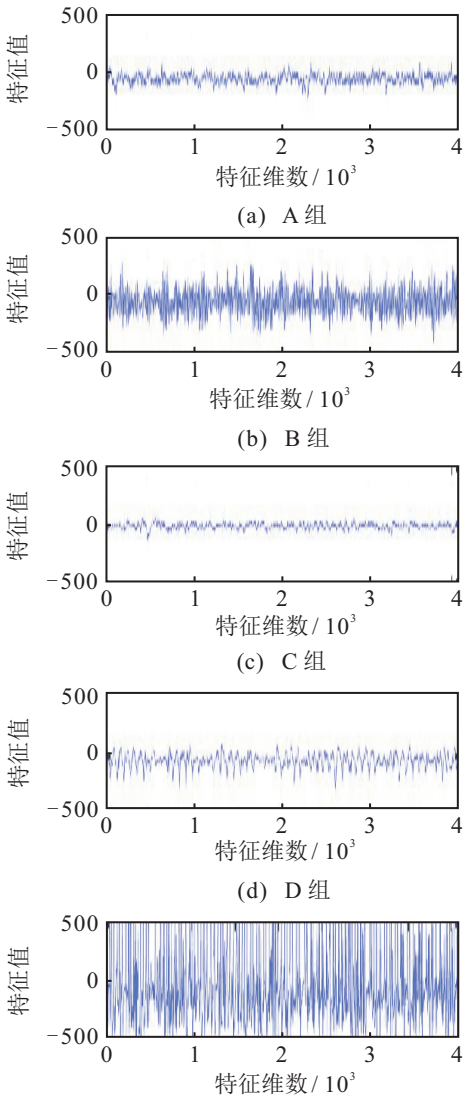


图4 EEG原始信号

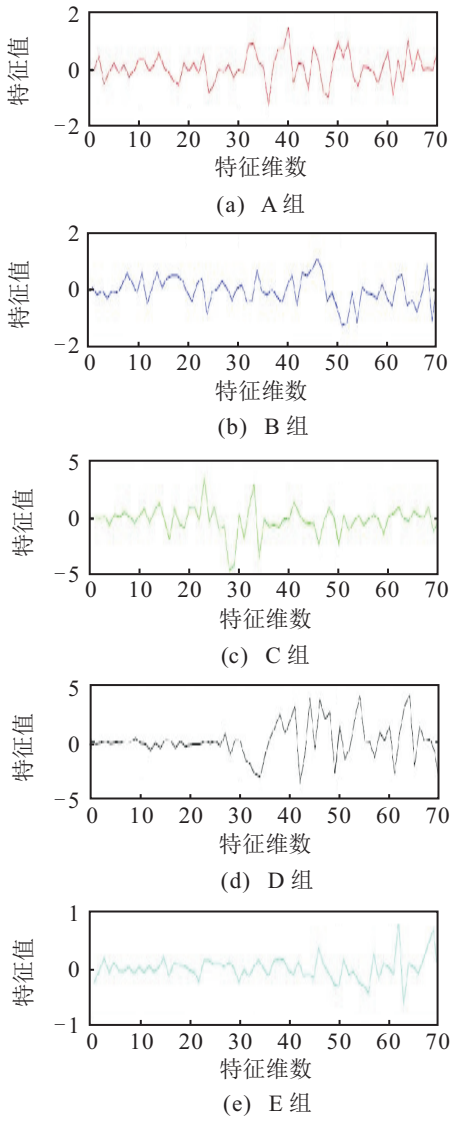


图 5　经K-PCA特征降维后的EEG信号

表 2　真实数据集详细配置

数据集	样本数	特征数	聚类数
EEG-D1	200	70	2
EEG-D2	200	70	2
EEG-D3	300	70	3
HWD1	4 000	7	2
HWD2	4 000	7	2
HWD3	6 000	7	3
VD1	180	13	2
VD2	180	13	2
VD3	270	13	3

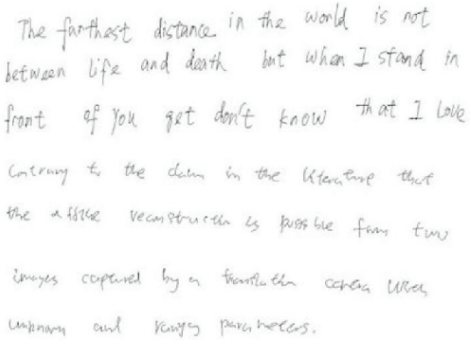


图 6　手写体部分数据展示

因此相应的手写数据风格也相互区别. 实验中分别将来自以下作者的手写体数据组成真实数据集: luan和taojing, liuchuankai和lulingling, chaowenting、fuyu和maying, 并将这3个数据集分别命名为HWD1、HWD2和HWD3. 另外, 参照文献[1], 从每一位作者的手写体数据中随机选取2 000个样本组成以上3个数据集. 3个真实数据集的详细配置如表2所示.

对于元音识别[1, 8], 表3列出了每个具体元音展示. 由此可知, 每个元音的发音各不相同, 不管男女, 同一个元音的发音相同, 意味着同一个元音的发音具有相同的风格. 实验中分别将元音数据集Vowel[1, 8]中标签为3和5, 标签为6和7, 标签为1、4和9对应的样本组成相应的真实数据集, 并分别命名为VD1、VD2和VD3. 3个真实数据集的详细配置如表2所示.

表 3　英文中的元音展示

vowel	word	vowel	word
æ	hoard	α:	hard
ʊ	hood	i:	heed
SD3	head	I	hid
SD4	heard	Λ	hud
æ	had	u:	who'd
SD6	hod		

2.3　结果与分析

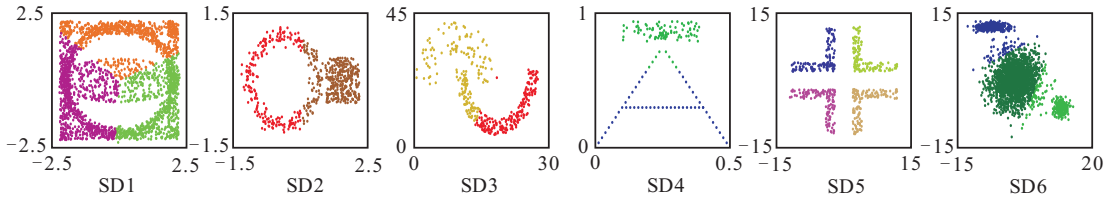
图7和表4分别为所有对比算法在人造数据集上的聚类结果. 表4中: 正确率(Acc)、F-measure和Rand Index(RI)为外部类型的聚类评价指标[9-10,25], Davies-Bouldin Index(DBI)为内部类型的聚类评价指标[26], "–"代表算法参数采用默认设置, "––"代表标准差小于10^{-4}. 最好的聚类正确率用黑体标出. 由图7和表4可得出以下结论:

1) 就聚类正确率而言, ISF-KPC在人造数据集SD1、SD2、SD3和SD5上取得最好的聚类结果, 尤其对于包含典型数据风格的数据集SD1、SD2和SD3, ISF-KPC能够取得优越的聚类性能. 就其他内外部聚类评价指标而言, ISF-KPC至少能够取得竞争性的聚类性能.

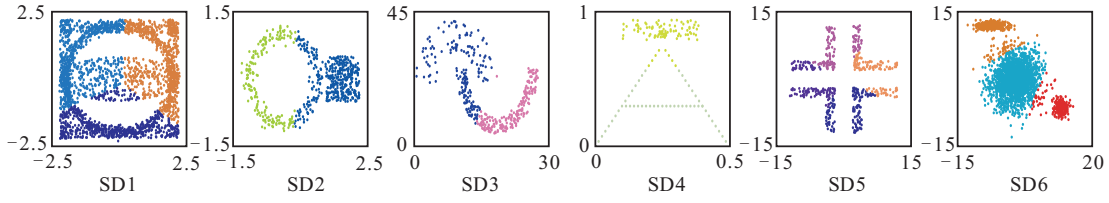
2) 对于其他对比算法, 由于能够识别任意形状的聚类, DBSCAN在人造数据集上能够取得较好的聚类结果.

3) 由于ISF-KPC考虑了数据双知识表达, 即特征增量学习和数据风格信息学习, ISF-KPC的聚类性能明显优于仅考虑数据风格信息的ISF-KPC_0.

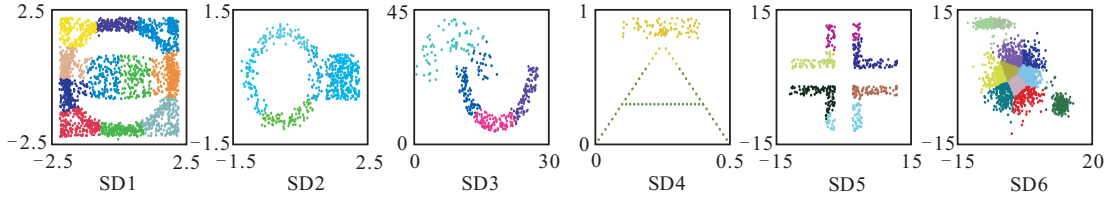
4) 将ISF-KPC、ISF-KPC_0分别与KPC及F-KPC比较, 相关实验结果有力地验证了挖掘数据风格信息并用于提高聚类性能的有效性.



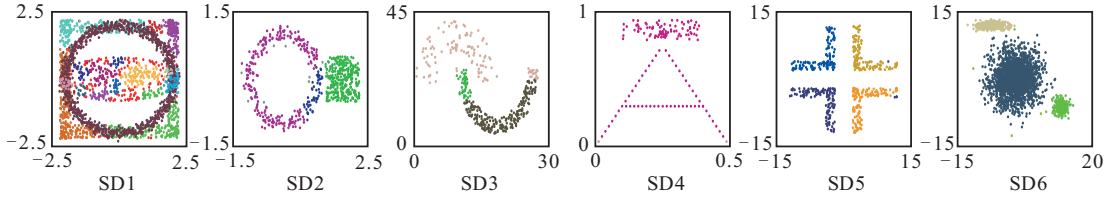
(a) FCM



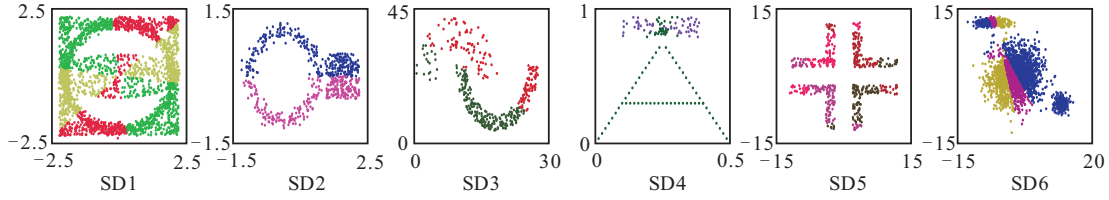
(b) *K*-medoids



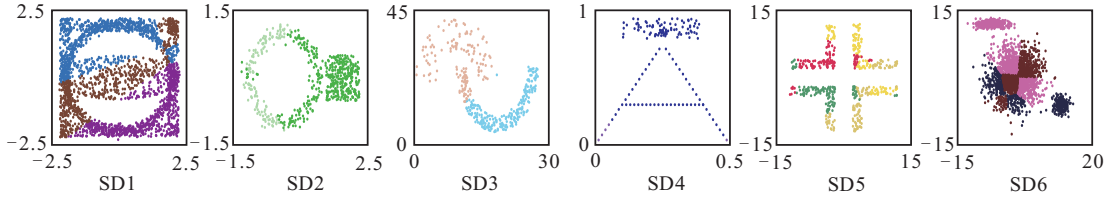
(c) AP



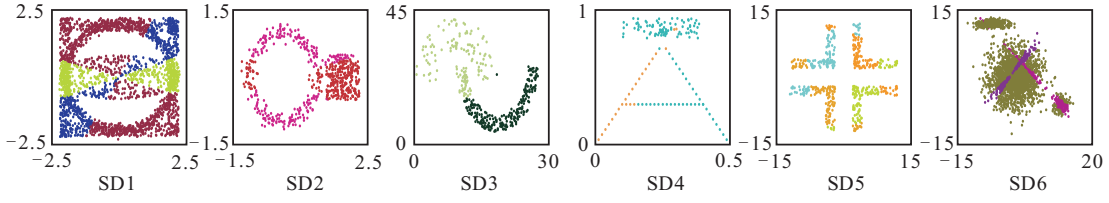
(d) DBSCAN



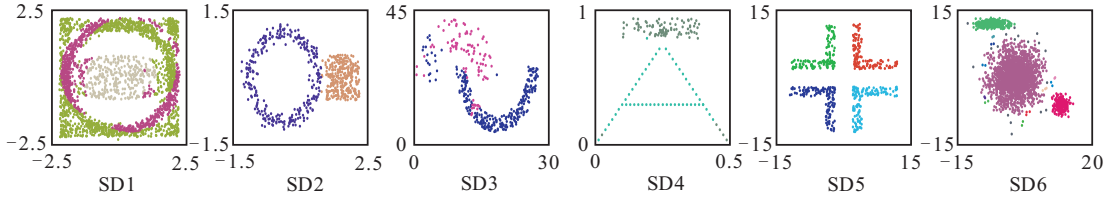
(e) KPC



(f) F-KPC



(g) ISF-KPC_0



(h) ISF-KPC

图 7 对比算法在人造数据集上的聚类结果视觉展示

表 4 所有对比算法在人造数据集上的详细聚类结果

Methods	SD1					SD2				
	Acc	F-measure	RI	DB	Opt	Acc	F-measure	RI	DB	Opt
FCM	0.349 9 (0.007 7)	0.348 7 (0.001 0)	0.546 3 (0.001 3)	0.008 8 (0.000 1)	–	0.848 3 (0)	0.753 2 −−	0.742 2 −−	0.007 8 (0)	–
K-medoids	0.350 3 (0.008 1)	0.350 5 (0.001 2)	0.545 4 (0.001 7)	0.008 7 (0.000 2)	$K=3$	0.805 0 (0)	0.707 3 (0)	0.685 5 (0)	0.007 8 (0)	$K=2$
AP	0.188 7 (0)	0.207 7 −−	0.639 8 −−	0.006 7 (0.000 3)	PR $=-82.12$	0.535 0 (0)	0.610 3 (0)	0.707 7 −−	0.007 4 −−	PR $=-27.31$
DBSCAN	0.559 0 (0)	0.708 2 −−	0.836 9 (0)	0.012 3 (0)	Eps $=0.20$ MinPts $=13$	0.900 0 (0)	0.910 0 (0)	0.917 6 (0)	0.009 5 (0)	Eps $=0.15$ MinPts $=6$
KPC	0.388 6 (0)	0.358 0 (0)	0.549 4 (0)	0.061 2 (0.030 5)	$K=3$	0.505 0 (0)	0.498 5 (0)	0.499 2 (0)	0.024 3 (0)	$K=2$
F-KPC	0.429 4 (0.010 2)	0.370 1 (0.006 1)	0.562 5 (0.003 3)	0.021 3 (0.005 1)	$K=3$	0.733 3 (0)	0.656 6 (0)	0.608 2 (0)	0.008 6 (0)	$K=2$
ISF-KPC_0	0.448 2 (0.024 5)	0.400 8 (0.048 7)	0.512 2 (0.053 4)	0.042 2 (0.052 0)	$\lambda=10^2$	0.798 3 (0.063 8)	0.765 4 (0.046 4)	0.722 9 (0.037 6)	0.045 2 (0.021 0)	$\lambda=10^{-1}$
ISF-KPC	**0.606 2** (0.071 2)	0.594 3 (0.049 3)	0.617 9 (0.090 1)	0.099 1 (0.021 0)	$\lambda=10$ $\sigma_1=10^2,\sigma_2=10$	**0.990 0** (0.003 3)	0.986 4 (0.004 4)	0.981 6 (0.003 2)	0.006 2 (0.000 1)	$\lambda=10^4$ $\sigma_1=10^2,\sigma_2=10$

Methods	SD3					SD4				
	Acc	F-measure	RI	DB	Opt	Acc	F-measure	RI	DB	Opt
FCM	0.774 8 (0)	0.685 8 −−	0.650 1 (0)	0.008 6 −−	–	0.946 7 −−	0.904 2 −−	0.898 3 −−	0.005 3 −−	–
K-medoids	0.766 8 (0)	0.677 8 (0)	0.641 4 (0)	0.008 6 (0)	$K=2$	0.946 7 −−	0.904 2 −−	0.898 3 −−	0.005 3 −−	$K=2$
AP	0.466 5 (0)	0.512 8 (0)	0.577 5 (0)	0.007 8 (0.000 2)	PR $=-4\,031.25$	0.946 7 (0)	0.904 2 (0)	0.898 3 (0)	0.005 3 (0)	PR $=-0.79$
DBSCKN	0.828 4 (0)	0.774 6 (0)	0.763 8 (0)	0.006 4 (0)	Eps $=2.15$ MinPts $=25$	**0.973 3** −−	0.980 2 (0)	0.980 0 (0)	0.005 1 (0)	Eps $=0.15$ MinPts $=9$
KPC	0.753 4 (0)	0.670 5 (0)	0.627 4 (0)	0.022 0 (0)	$K=2$	0.773 3 (0)	0.662 5 (0)	0.647 1 (0)	0.011 6 (0)	$K=2$
F-KPC	0.864 2 (0.002 5)	0.792 8 (0.003 6)	0.764 6 (0.003 7)	0.008 7 −−	$K=2$	0.568 9 (0.056 7)	0.645 5 (0.024 8)	0.512 7 (0.019 4)	0.015 5 (0.008 0)	$K=2$
ISF-KPC_0	0.893 9 (0.024 7)	0.698 7 (0.007 8)	0.674 5 (0.024 7)	0.010 1 (0.001 5)	$\lambda=10$	0.828 3 (0.027 9)	0.843 0 (0.033 4)	0.809 9 (0.009 5)	0.095 4 (0.046 7)	$\lambda=1$
ISF-KPC	**0.900 8** (0.016 9)	0.904 6 (0.085 8)	0.914 1 (0.089 8)	0.030 1 (0.021 7)	$\lambda=1$ $\sigma_1=10^2,\sigma_2=10^2$	0.960 0 (0.085 1)	0.971 6 (0.084 1)	0.972 5 (0.089 3)	0.004 9 (0.009 8)	$\lambda=1$ $\sigma_1=10^2,\sigma_2=10^2$

Methods	SD5					SD6				
	Acc	F-measure	RI	DB	Opt	Acc	F-measure	RI	DB	Opt
FCM	0.900 4 (0.161 4)	0.872 2 (0.198 5)	0.936 4 (0.098 8)	0.008 1 (0.000 3)	–	0.946 5 −−	0.923 6 −−	0.910 2 −−	0.006 2 (0)	–
K-medoids	0.684 5 (0)	0.565 6 (0.000 1)	0.783 9 (0.000 1)	0.008 2 (0.000 4)	$K=4$	0.963 8 −−	0.947 7 (0)	0.937 3 −−	0.005 3 (0.000 6)	$K=3$
AP	0.757 9 (0)	0.761 2 −−	0.899 2 −−	0.007 8 (0.000 5)	PR $=-1\,235.18$	0.300 8 (0)	0.269 6 (0)	0.474 7 −−	0.008 2 (0.000 1)	PR $=-1\,197.86$
DBSCAN	0.992 1 (0)	0.992 1 (0)	0.996 1 (0)	0.005 7 (0)	Eps $=1.00$ MinPts $=2$	**0.998 1** (0)	0.997 6 (0)	0.997 0 (0)	0.002 7 −−	Eps $=2.00$ MinPts $=3$
KPC	0.514 2 (0)	0.505 6 (0)	0.683 9 (0)	0.012 6 (0)	$K=4$	0.381 5 (0)	0.447 8 (0)	0.460 0 (0)	0.044 7 (0)	$K=3$
F-KPC	0.676 6 (0.044 9)	0.534 0 (0.058 1)	0.768 1 (0.028 5)	0.014 8 (0.000 7)	$K=4$	0.498 5 (0.026 6)	0.501 3 (0.022 1)	0.495 6 (0.007 8)	0.025 0 (0.004 4)	$K=3$
ISF-KPC_0	0.720 2 (0.080 4)	0.532 4 (0.029 0)	0.767 5 (0.014 4)	0.100 5 (0.001 9)	$\lambda=10^{-2}$	0.722 6 (0.094 3)	0.667 8 (0.078 9)	0.712 4 (0.032 8)	0.107 8 (0.051 5)	$\lambda=1$
ISF-KPC	**1** (0)	1 (0)	1 (0)	0.008 0 (0.000 4)	$\lambda=10^4$ $\sigma=10^2$	0.984 6 −−	0.981 2 (0.000 1)	0.977 1 (0.000 3)	0.003 5 −−	$\lambda=10^{-3}$ $\sigma_1=10^2,\sigma_2=10,\sigma_3=10$

表 5 所有对比算法在真实数据集上的详细聚类结果

Methods	EEG-D1					EEG-D2				
	Acc	F-measure	RI	DB	Opt	Acc	F-measure	RI	DB	Opt
FCM	0.522 5 (0.010 5)	0.499 0 (0.003 0)	0.498 7 (0.001 0)	0.058 4 (0.001 5)	–	0.526 5 0.019 5	0.497 8 0.003 1	0.499 7 0.002 8	0.069 9 0.001 9	–
K-medoids	0.585 0 (0)	0.510 8 (0)	0.512 0 ––	0.066 6 ––	$K=2$	0.510 0 ––	0.660 0 (0)	0.497 7 ––	0.006 0 (0)	$K=2$
AP	0.500 6 (0)	0.664 4 (––)	0.497 5 (––)	0.073 2 (0.000 6)	PR $=-1\,262.13$	0.492 5 (0)	0.603 4 ––	0.486 7 ––	0.030 5 (0.000 2)	PR $=-2\,553.33$
DBSCAN	0.575 0 (0)	0.638 3 (0)	0.508 8 (0)	0.068 1 (0)	Eps $=3.00$ MinPts $=15$	0.590 0 (0)	0.630 7 (0)	0.513 8 (0)	0.033 7 (0)	Eps $=3.95$ MinPts $=20$
KPC	0.515 0 (0)	0.663 0 (0)	0.530 2 (0)	0.095 6 (0)	$K=2$	0.540 0 (0)	0.632 9 ––	0.521 1 (0)	0.097 4 ––	$K=2$
F-KPC	0.536 7 (0.045 5)	0.548 8 (0.060 5)	0.529 5 (0.064 2)	0.097 9 (0.004 4)	$K=2$	0.588 3 (0.086 1)	0.656 2 (0.090 2)	0.579 5 (0.106 9)	0.089 7 (0.011 3)	$K=2$
ISF-KPC_0	0.526 0 (0.025 4)	0.498 6 (0.004 0)	0.500 1 (0.004 3)	0.113 4 (0.001 4)	$\lambda=10^2$	0.519 0 (0.008 6)	0.496 5 (0.001 2)	0.498 4 (0.000 8)	0.118 6 (0.007 2)	$\lambda=10^2$
ISF-KPC	**0.632 5** (0.010 0)	0.661 8 (0.000 8)	0.528 9 (0.001 0)	0.042 2 (0.008 5)	$\lambda=1$ $\sigma_1=10^2,\sigma_2=10$	**0.652 1** (0.002 5)	0.683 9 (0.068 5)	0.563 8 (0.000 1)	0.060 3 (0.027 4)	$\lambda=10^{-2}$ $\sigma_1=10,\sigma_2=10$

Methods	ECG-D3					HWD1				
	Acc	F-measure	RI	DB	Opt	Acc	F-measure	RI	DB	Opt
FCM	0.450 0 ––	0.427 7 ––	0.526 5 (0)	0.050 2 0.0022	–	0.554 2 –	0.512 2 (0)	0.505 8 ––	0.007 2 ––	–
K-medoids	0.369 0 (0.003 0)	0.404 0 (0.000 1)	0.496 6 (0.000 6)	0.043 6 (0.000 3)	$K=3$	0.565 9 (0.018 4)	0.516 9 (0.008 7)	0.509 2 (0.004 3)	0.007 2 (0.000 2)	$K=2$
AP	0.356 4 ––	0.482 2 ––	0.337 9 (0)	0.082 5 (0.000 2)	PR $=-2\,684.71$	0.281 3 (0)	0.382 2 ––	0.618 2 ––	0.004 8 (0.000 2)	PR $=10^9$
DBSCAN	0.560 0 (0)	0.520 8 (0)	0.557 8 (0)	0.006 9 (0)	Eps $=2.55$ MinPts $=5$	0.509 3 (0)	0.656 0 (0)	0.500 6 (0)	0.000 1 (0)	EPS $=1.00$ MinPts $=11$
KPC	0.366 7 (0)	0.492 3 (0)	0.362 6 (0)	0.077 3 (0)	$K=3$	0.480 2 (0)	0.634 8 ––	0.509 4 (0)	0.105 7 (0)	$K=2$
F-KPC	0.352 2 (0.004 2)	0.487 8 (0.002 0)	0.358 9 (0.003 5)	0.020 8 (0.002 8)	$K=3$	0.667 4 (0.136 4)	0.664 3 (0.001 7)	0.593 1 (0.076 2)	0.023 3 (0.011 3)	$K=2$
ISF-KPC_0	0.546 7 (0.015 6)	0.506 5 (0.024 5)	0.642 4 (0.012 5)	0.102 3 (0.014 7)	$\lambda=10^2$	0.679 7 (0.115 6)	0.577 1 (0.056 8)	0.601 4 (0.058 8)	0.050 3 (0.020 4)	$\lambda=10^6$
ISF-KPC	**0.619 1** (0.016 7)	0.554 2 (0.006 2)	0.564 5 ––	0.011 9 (0.004 1)	$\lambda=10^{-1}$ $\sigma_1=10$	**0.808 1** (0.005 9)	0.702 8 (0.009 8)	0.689 9 (0.007 2)	0.031 4 (0.000 2)	$\lambda=10^9$ $\sigma=10^2,\sigma_2=10^{-1}$

Methods	HWD2					HWD3				
	Acc	F-measure	RI	DB	Opt	Acc	F-measure	RI	DB	Opt
FCM	0.503 8 (0)	0.5147 (0)	0.499 9 (0)	0.009 0 ––	–	0.377 3 ––	0.339 9 ––	0.556 2 ––	0.005 3 (0.000 8)	–
K-medoids	0.539 7 (0.051 1)	0.523 8 (0.002 6)	0.508 3 (0.012 7)	0.009 0 (0.000 4)	$K=2$	0.384 5 (0.004 0)	0.344 0 (0.001 6)	0.557 0 (0.001 1)	0.005 5 (0.000 7)	$K=3$
AP	0.347 2 ––	0.429 1 ––	0.636 7 ––	0.005 3 (0.000 2)	PR $=-10^9$	0.199 0 ––	0.205 7 ––	0.632 0 ––	0.006 6 (0.000 3)	PR $=-10^9$
DBSCAN	0.505 8 (0)	0.652 8 (0)	0.500 6 (0)	0.000 3 (0)	Eps $=2.25$ MinPts $=10$	0.347 3 (0)	0.487 7 (0)	0.368 3 (0)	0.001 2 (0)	Eps $=3.00$ MinPts $=4$
KPC	0.528 7 (0)	0.636 1 ––	0.512 7 (0)	0.110 4 ––	$K=2$	0.386 2 (0)	0.485 3 (0)	0.362 0 ––	0.056 8 (0)	$K=3$
F-KPC	0.699 0 (0)	0.628 2 (0)	0.623 9 (0)	0.028 4 (0)	$K=2$	0.592 7 (0.005 5)	0.667 9 (0.009 1)	0.693 8 (0.010 0)	0.102 8 (0.007 4)	$K=3$
ISF-KPC_0	0.659 1 (0.104 0)	0.579 5 (0.064 7)	0.572 2 (0.065 7)	0.052 2 (0.029 9)	$\lambda=10^6$	0.610 0 (0.062 3)	0.661 6 (0.026 7)	0.710 9 (0.031 4)	0.105 3 (0.034 7)	$\lambda=10^6$
ISF-KPC	**0.713 8** (0.059 5)	0.601 9 (0.054 3)	0.598 4 (0.050 9)	0.034 8 (0.008 2)	$\lambda=10^7$ $\sigma=10^2$	**0.618 5** (0.049 9)	0.670 0 (0.052 3)	0.697 7 (0.071 8)	0.329 7 (0.121 3)	$\lambda=10^7$ $\sigma=10^3$

表5 (续)

Methods	VD1					VD2				
	Acc	F-measure	RI	DB	Opt	Acc	F-measure	RI	DB	Opt
FCM	0.5111 (0)	0.4951 −−	0.4975 (0)	0.0070 −−	−	0.5000 (0)	0.4955 −−	0.4972 −−	0.0076 (0)	−
K-medoids	0.5056 (0)	0.4952 −−	0.4973 (0)	0.0070 (0)	$K=2$	0.5000 (0)	0.4955 −−	0.4972 −−	0.0078 −−	$K=2$
AP	0.5167 (0)	0.5085 −−	0.4978 (0)	0.0076 −−	PR $=-788.07$	0.5056 (0)	0.4952 −−	0.4973 (0)	0.0070 −−	PR $=-655.85$
DBSCAN	0.4944 (0)	0.4768 −−	0.4988 (0)	0.0066 −−	Eps $=2.50$ MinPts $=20$	0.5611 (0)	0.5810 (0)	0.5047 (0)	0.0072 (0)	Eps $=2.15$ MinPts $=19$
KPC	0.5078 (0)	0.5101 −−	0.4984 (0)	0.0450 (0)	$K=2$	0.5278 −−	0.5636 (0)	0.5008 (0)	0.1195 (0)	$K=2$
F-KPC	0.5037 (0.0224)	0.5422 (0.0362)	0.5098 (0.0087)	0.0762 (0.0321)	$K=2$	0.5300 (0.0091)	0.5229 (0.0106)	0.5326 (0.0055)	0.0664 (0.0187)	$K=2$
ISF-KPC_0	0.5300 (0.0152)	0.5286 (0.0382)	0.4995 (0.0017)	0.0848 (0.0831)	$\lambda=10$	0.5556 (0.0256)	0.5249 (0.0390)	0.5047 (0.0049)	0.0621 (0)	$\lambda=10$
ISF-KPC	**0.6220** (0.0139)	0.5775 (0.0731)	0.5021 (0.0026)	0.0471 (0.0226)	$\lambda=10^{-2}$ $\sigma_1=10^2,\sigma_2=10$	**0.6167** (0.0056)	0.5803 (0.0041)	0.5195 (0.0007)	0.0549 (0.0362)	$\lambda=10^{-2}$ $\sigma=10^{-1},\sigma_2=1$

Methods	VD3				
	Acc	F-measure	RI	DB	Opt
FCM	0.3556 (0)	0.3283 (0)	0.5549 −−	0.0096 (0.0014)	−
K-medoids	0.3626 (0.0164)	0.3394 (0.0021)	0.5494 (0.0041)	0.0104 (0.0011)	$K=3$
AP	0.3333 −−	0.3953 −−	0.4974 −−	0.0084 −−	PR $=-859.61$
DBSCAN	0.4000 (0)	0.4386 (0)	0.7621 −−	0.0081 (0)	Eps $=2.50$ MinPts $=5$
KPC	0.3733 (0)	0.3601 (0)	0.5326 (0)	0.0691 (0)	$K=3$
F-KPC	0.3963 (0.0246)	0.3683 (0.0321)	0.5346 (0.0968)	0.0535 (0.0161)	$K=3$
ISF-KPC_0	0.3815 (0.0221)	0.3523 (0.0162)	0.5457 (0.0105)	0.1100 (0.0286)	$\lambda=10$
ISF-KPC	**0.4556** (0.0500)	0.4643 (0.0383)	0.5658 (0.0746)	0.0367 (0.0108)	$\lambda=10^{-1}$ $\sigma_1=10,\sigma_2=10,\sigma_3=1$

5) 在原输入特征基础上的增量学习次数只需1 ∼ 3次,增量学习过程简单且易于执行.

表5详细列出了所有对比算法在真实数据集上的聚类结果,且每一个真实数据集均包含典型的数据风格. 由表5可见,就聚类正确率而言,ISF-KPC在所有真实数据集上取得最好的聚类结果,就其他内外部聚类评价指标而言,ISF-KPC在多数情况下能够取得最好的聚类性能. 与其他对比算法相比(除ISF-KPC_0之外),由于考虑了数据风格信息,ISF-KPC在多数情况下至少能够取得竞争性的聚类性能.

为了进一步比较ISF-KPC与其他对比算法的区别,利用文献[27]的统计测试方法对所有对比算法的聚类性能进行分析.

该统计测试方法主要包含F_F和CD两个关键值[27]. F_F用于确定该统计测试方法的空假设(即假设所有对比算法具有相同的聚类性能)是否被否定;CD用于验证对比算法之间的聚类性能是否存在显著区别. 由表5列出的聚类正确率对所有对比算法

进行排序,如对于EEG-D2,因ISF-KPC取得了最好的聚类结果,其排名为1,即rank $=1$. 依此类推,可得出所有对比算法在所有真实数据集上的排序, 如表6所示. 根据对比算法排序值、F-分布及其自由度(N_c-1)、$(N_c-1)(N_d-1)$可确定F_F值. 只要满足$F_F > F((N_c-1),(N_d-1))$,该统计测试方法的空假设即被否定. 其中N_c代表所有对比算法个数,N_d代表真实数据集个数,$F(\cdot,\cdot)$根据F-分布表可查出[27]. 基于空假设被否定,可进一步计算CD值. 其中$\mathrm{CD} = q_\alpha\sqrt{\dfrac{N_c(N_c+1)}{6N_d}}$($\alpha$为显著性度,其值取文献[27]推荐值$\alpha=0.05$,$q_\alpha$可从文献[27]查出). 任意两个对比算法平均排序差值的绝对值大于该CD值即表明这两个对比算法的聚类性能之间存在显著区别.

表6 所有对比算法在真实数据集上的排序

methods	FCM	K-medoids	AP	DBSCAN	KPC	F-KPC	ISF-KPC_0	ISF-KPC
EEG-D1	6	2	8	3	7	4	5	1
EEG-D2	5	7	8	2	4	3	6	1
EEG-D3	4	5	7	2	6	8	3	1
HWD1	5	4	8	6	7	3	2	1
HWD2	7	4	8	6	5	2	3	1
HWD3	6	5	8	7	4	3	1	1
VD1	4	6	3	8	5	7	2	1
VD2	7	7	6	2	5	4	3	1
VD3	7	6	8	2	5	3	4	1
average rank(AR)	5.67	5.11	7.11	4.22	5.33	4.11	3.33	1.00

由表5和表6可知,$F(7,56)\approx 2.18$、$F_F\approx 8.87$,因此该统计测试方法的空假设被否定. 由给出的显著性度$\alpha=0.05$可计算$\mathrm{CD}\approx 3.12$. 根据统计测试结果可得出以下结论:

1) 由于ISF-KPC与其他对比算法(除ISF-KPC_0与F-KPC外)之间的平均排序差值都大于$\mathrm{CD}\approx 3.12$,ISF-KPC与其他对比算法之间存在显著区别. 另外,由于ISF-KPC考虑了双知识表达,其聚类性能优于ISF-KPC_0和F-KPC.

2) 由于挖掘并利用了数据风格信息约束聚类行为,ISF-KPC_0的聚类性能优于其他对比算法(除ISF-KPC外),特别地,ISF-KPC_0与AP算法之间存在显著区别.

3) 结合1)和2),在真实数据集上的统计测试结果充分表明了数据双知识表达确实能够提高聚类算法性能,尤其基于数据风格信息的聚类算法,其聚类性能能够显著区别于典型聚类算法.

3 结 论

由于数据集中的每一类数据都潜在或明显地具有独特的数据风格,这些风格信息显著区别于数据物理特征. 如何挖掘隐藏的数据风格信息并用于聚类分析是一个值得研究的课题. 本文利用风格矩阵迭代地捕捉数据集中每个聚类的数据风格信息,这些风格矩阵相互区别. 另外,沿着高斯型SVM将输入特征从低维空间映射到高维空间的思想,本文在原输入特征空间中增量式地对原输入特征进行扩维. 通过增量学习和数据风格矩阵对数据进行双知识表达以强化聚类性能. 人造数据集尤其真实数据集上的实验结果验证了ISF-KPC中双知识表达的有效性. 真实数据集上的统计测试结果表明ISF-KPC与典型聚类方法之间存在显著区别. 未来的研究会重点关注如何将ISF-KPC扩展到回归模型,进一步研究本文双知识表达并将ISF-KPC推广到深度学习.

参考文献(References)

[1] Huang K Z, Jiang H C, Zhang X Y. Field support vector machines[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, 1(6): 454-463.

[2] Jiang S H, Shao M, Jia C C, et al. Learning consensus representation for weak style classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2906-2919.

[3] Veeramachaneni S, Nagy G. Analytical results on style-constrained Bayesian classification of pattern fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(7): 1280-1285.

[4] Sarkar P, Nagy G. Style consistent classification of isogenous patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(1): 88-98.

[5] Jiang Y Z, Deng Z H, Wang S T, et al. Recognition of epileptic EEG signals using a novel multiview TSK fuzzy system[J]. IEEE Transactions on Fuzzy Systems, 2017,

25(1): 3-20.

[6] Xie L X, Deng Z H, Wang S T, et al. Generalized hidden-mapping transductive transfer learning for recognition of epileptic electroencephalogram signals[J]. IEEE Transactions on Cybernetics, DOI: 10.1109/TCYB.2018.2821764.

[7] Chinese Academy of Sciences Institute of Automation (CASIA). Handwriting database[DB/OL]. [2018-10-07]. http://biometrics.idealtest.org/.

[8] Zhang X Y, Huang K Z, Liu C L. Pattern field classification with style normalized transformation[C]. Proceedings of the 22th International Joint Conference on Artificial Intelligence. Spain: IEEE, 2011: 1621-1626.

[9] 陈爱国, 王士同. 基于多代表的大规模数据模糊聚类算法[J]. 控制与决策, 2016, 31(12): 2122-2130.
(Chen A G, Wang S T. Fuzzy clustering algorithm based on multiple medoids for large-scale data[J]. Control and Decision, 2016, 31(12): 2122-2130.)

[10] 乔颖, 王士同, 杭文龙. 大规模数据集引力同步聚类[J]. 控制与决策, 2017, 32(6): 1075-1083.
(Qiao Y, Wang S T, Hang W L. Clustering by gravitational synchronization on large scale dataset[J]. Control and Decision, 2017, 32(6): 1075-1083.)

[11] Bradley P S, Mangasarian O L. k-plane clustering[J]. Journal of Global Optimization, 2000, 16(1): 23-32.

[12] Shao Y H, Bai L, Wang Z, et al. Proximal plane clustering via eigenvalues[J]. Procedia Computer Science, 2013, 17: 41-47.

[13] Jayadeva, Khemchandani R, Chandra S, et al. Twin support vector machines for pattern classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905-910.

[14] Wang Z, Shao Y H, Bai L, et al. Twin support vector machine for clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 26(10): 2583-2588.

[15] Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.

[16] Bai X, Chen Z, Zhang Y, et al. Infrared ship target segmentation based on spatial information improved FCM[J]. IEEE Transactions on Cybernetics, 2016, 46(12): 3259-3271.

[17] Anand S, Mittal S, Tuzel O, et al. Semi-supervised kernel mean shift clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(6): 1201-1215.

[18] Fang X Z, Wong W K, Teng S H, et al. Flexible affinity matrix learning for unsupervised and semisupervised classification[J]. IEEE Transactions on Neural Networks and Learning Systems, DOI: 10.1109/TNNLS.2018.2861839.

[19] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315: 972-976.

[20] Arzeno N M, Vikalo H. Semi-supervised affinity propagation with soft instance-level constraints[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(5): 1041-1052.

[21] Sakai T, Tamura K, Kitakami H. Density-based adaptive spatial clustering algorithm for identifying local high-density areas in georeferenced documents[C]. Proceedings IEEE International Conference System, Man, and Cybernetics. New York: IEEE, 2014: 513-518.

[22] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. Proceedings of ACM Conference on Knowledge Discovery and Data Mining. New York: ACM, 1996: 226-231.

[23] Jiang Y Z, Deng Z H, Wang S T, et al. Realizing two-view TSK fuzzy classification system by using collaborative learning[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 47(1): 145-160.

[24] Hou J, Gao H J, Li X L. DSets-DBSCAN: A parameter-free clustering algorithm[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3182-3193.

[25] Wang Y T, Chen L H, Mei J P. Incremental fuzzy clustering with multiple medoids for large data[J]. IEEE Transactions on Fuzzy Systems, 2014, 22(6): 1557-1568.

[26] Davies D L, Bouldin D W. A cluster separation measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, PAMI-1(2): 224-227.

[27] Demsar J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006, 7: 1-30.

作者简介

顾苏杭 (1989−), 男, 博士生, 从事人工智能与模式识别、机器学习的研究, E-mail: gusuhang09@163.com;

王士同 (1964−), 男, 教授, 博士生导师, 从事人工智能与模式识别、机器学习、深度学习等研究, E-mail: wxwangst@yahoo.com.cn.

(责任编辑: 郑晓蕾)