结构α-熵的加权高斯混合模型的子空间聚类

李 凯[1,2],张可心[1]
(1. 河北大学网络空间安全与计算机学院,河北保定071002; 2. 河北省机器视觉工程研究中心,河北保定071002)

摘 要: 利用信息熵或模糊熵确定子空间聚类中每个簇的不同特征,较好地解决了高维数据的子空间聚类. 为了进一步提高聚类算法的性能,将权向量的负结构α-熵引入到高斯混合模型中,获得了结构α-熵的加权高斯混合的子空间聚类模型,提出了结构α-熵的加权高斯混合模型的子空间聚类算法SEWMM(Structural α-Entropy Weighting Mixture Model),该算法不仅可以发现高维数据空间中位于不同子空间的簇,而且能够获得子空间中具有不同形状体积的簇. 同时,进一步分析了算法的收敛性与时间复杂性. 通过选取UCI(University of California, Irvine)标准数据集及图像数据集,对提出的算法SEWMM进行了实验,并与一些典型的聚类算法进行了比较,表明了提出的算法在总体性能上具有一定的提升.

关键词: 模糊熵; 结构α-熵; 特征加权; 高斯混合模型; 高维数据; 子空间聚类
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112(2022)03-0718-08
电子学报URL: http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20201146

Structural α-Entropy Weighting Gaussian Mixture Model for Subspace Clustering

LI Kai[1,2], ZHANG Ke-xin[1]
(1. *School of Cyber Security and Computer, Hebei University, Baoding, Hebei* 071000, *China*;
2. *Hebei Machine Vision Engineering Research Center, Baoding, Hebei* 071000, *China*)

Abstract: Using information entropy or fuzzy entropy to determine the different features of each cluster for subspace clustering, subspace clustering for high dimensional data is solved very well. For further improving performance of clustering algorithm, negative structural α-entropy with weight vector is introduced into the Gaussian mixture model to obtain a structural α-entropy weighting mixture model of subspace clustering. Based on this, the structural α-entropy weighting mixture model subspace clustering algorithm(SEWMM) is derived theoretically, which can not only discover clusters in different subspaces in high dimensional data space, but also can discover clusters with various shape volumes in subspaces. And convergence and time complexity of algorithm are further analyzed. In the experiment, compared with some representative algorithms, the proposed algorithm SEWMM is tested on UCI(University of California, Irvine) standard data sets and image data sets. It shows the proposed algorithm has a certain improvement in the overall performance.

Key words: fuzzy entropy; structural α-entropy; feature weighting; Gaussian mixture model; high-dimensional data; subspace clustering

1 引言

近年来,随着信息技术、计算机网络与人工智能技术的不断发展,在实际问题中产生了大量高维数据,如何对这些数据进行聚类却遇到了很大的挑战. 尽管人们提出了一些不同的聚类算法,但由于高维数据的特殊性,使得传统聚类算法并不能直接应用于这些数据,为此,人们对高维数据的聚类进行了研究,其中方法之一是维数约简技术,该方法主要由降维和聚类两个独立的过程构成,且每个过程对聚类结果将会产生一定的影响,也就是说,利用该方法获得的嵌入特征对于聚类来说并不一定是最优的,且使得具有区分能力的信息丢失. 关于此问题,一些学者曾指出,若使用前几个

收稿日期:2020-10-18;修回日期:2021-03-30;责任编辑:覃怀银
基金项目:河北省自然科学基金(No.F2018201060)

主成分对数据聚类,则将破坏原数据的聚类结构[1]. 之后,人们提出了变量选择方法,较好地解决了降维和聚类过程独立的问题. 最近,Zhao等人[2]提出了基于正则项的高斯混合模型子空间聚类,此方法将聚类和寻找簇的子空间同时进行处理,进一步提高了子空间聚类的性能. Peng等人[3]为了获取不同形状与大小的子空间,将权向量的负信息熵引入到扩展的高斯混合模型中,提出了一种基于熵的高斯混合模型的子空间聚类方法,然而该方法只考虑了信息熵. 我们知道,信息熵是结构α-熵[4]的特例,特别是,对于结构α-熵中参数α的不同取值,则其对应于不同的熵,例如,当α趋近于1时,结构α-熵变为信息熵;当α=2时,则结构α-熵为平方熵. 另外,当α=2时,基于结构α-熵的最小化准则等于最近邻方法的错误概率. 以上表明,将基于权向量的信息熵作为惩罚项并不总是获得较好的聚类结果. 针对结构α-熵,Li等人[5]提出了最小熵的聚类方法,并将其应用于基因表达分析中,Nicholas等人[6]提出了基于结构α-熵度量的点集配准,并且通过结构α-熵中参数控制求解的类型.

为此,本文通过扩展信息熵,研究了结构α-熵的加权高斯混合模型的子空间聚类.

2　相关工作

子空间聚类是指将数据划分为不同的子空间或簇,并且寻求数据的多个簇结构. 通常,子空间聚类可以分为四种类型,下面对其进行简介. 基于统计的方法将数据视为独立抽取且服从高斯混合模型的样本,子空间聚类问题被转化为模型参数估计,可使用EM(Expectation Maximization)算法或最大似然方法估计相应的参数[3,7]. 由于高斯混合模型具有较强的概率解释以及对噪声鲁棒性等优点,因此,该方法获得了较广泛的应用[8],然而,对于传统的高斯混合模型,当面对高维数据的聚类时,由于需要估计大量参数,以及维数的逐渐增大,则极大似然估计问题很快成为不适用的方法. 为此,人们通过修改目标函数或对协方差矩阵加以约束,提出了基于惩罚项的高斯混合模型的子空间聚类[2]. 基于代数的方法包括两种形式,一种是矩阵分解方法[9,10],通过对数据矩阵分解达到数据分割的目的,其缺陷是对数据噪声和异常数据较为敏感;另一种是广义主成分分析法[11],它主要使用多项式拟合数据,但由于数据中存在噪声,因此,该方法的实现较为困难,且对高维数据的处理代价较高. 基于迭代的方法将样本分配给子空间,并利用迭代技术更新子空间的簇,从而实现对数据的划分[13,14]. 此类方法可划分为硬子空间聚类和软子空间聚类. 硬子空间聚类通过使用启发式准则,采取自顶向下或自底向上的搜索策略,从所有特征中寻找簇所在的子空间. 软子空间聚类通过分配特征权值确定簇所在的子空间,包括模糊加权子空间聚类[15~17]和熵加权子空间聚类[3,18,19]. 基于谱聚类的方法是将谱聚类算法作为框架,通过学习一个亲和矩阵找到数据的低维嵌入,并使用传统的聚类算法对低维数据聚类的一种方法. 亲和矩阵可以使用高斯核、局部信息或全局信息进行构造[20,21]. 最近,一些学者提出了对角为分块矩阵表示的子空间聚类以及复杂噪声下的子空间聚类,统一了现有的一些子空间聚类方法[22,23].

3　结构α-熵的加权高斯混合子空间聚类模型

假设样本集为$X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,其中$\boldsymbol{x}_k\in R^p(k=1,2,\cdots,n)$,且$\boldsymbol{x}_k$是由$c$个高斯分量组成的混合密度中抽取的样本,概率密度函数如式(1)所示.

$$F(\boldsymbol{x}_k;\ \Theta_0)=\sum_{i=1}^{c}\beta_i f(\boldsymbol{x}_k|\boldsymbol{V}_i,\boldsymbol{\Sigma}_i) \tag{1}$$

其中β_i、$\boldsymbol{V}_i$和$\boldsymbol{\Sigma}_i$分别为第i个分量的混合系数、均值向量和协方差矩阵;$f(\boldsymbol{x}_k|\boldsymbol{V}_i,\boldsymbol{\Sigma}_i)$为第$i$个分量的高斯密度函数. 为了解决高维数据的子空间聚类,Peng等人[3]对高斯分量的协方差矩阵特例化,如式(2)所示.

$$\boldsymbol{\Sigma}_i=\sigma_i{}^2\cdot\mathrm{diag}(w_{i1}^{-1},w_{i2}^{-1},\cdots,w_{ip}^{-1}) \tag{2}$$

其中$\mathrm{diag}(a_1,a_2,\cdots,a_p)$表示对角元素分别为$a_1,a_2,\cdots,a_p$的矩阵,$w_{ij}$表示特征$j$对簇$i$的相关性,且$\boldsymbol{w}_i=(w_{i1},w_{i2},\cdots,w_{ip})$是局部权向量,满足$0\le w_{ij}\le 1,\sum_{j=1}^{p}w_{ij}=1,\forall 1\le i\le c$;$\sigma_i{}^2$是簇$i$的局部方差. 利用式(2)后,则第$i$个高斯分量的概率密度函数变为

$$f(\boldsymbol{x}_k|\boldsymbol{V}_i,\sigma_i^2,\boldsymbol{w}_i)=\Big(\prod_{j=1}^{p}\sqrt{\frac{w_{ij}}{2\pi\sigma_i^2}}\Big)\exp\Big(-\frac{1}{2\sigma_i^2}\sum_{j=1}^{p}w_{ij}(V_{ij}-x_{kj})^2\Big) \tag{3}$$

为方便起见,令$\Theta=\{(\beta_i,\boldsymbol{V}_i,\sigma_i{}^2,\boldsymbol{w}_i)|1\le i\le c\}$. 为了估计$\Theta$中的参数,通过最小化KL散度[3],从而获得式(4):

$$G(\boldsymbol{u}_k,\beta_i,\boldsymbol{V}_i,\sigma_i^2,\boldsymbol{w}_i)=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\Big[-\log\beta_i+\sum_{j=1}^{p}\Big(\frac{w_{ij}}{2\sigma_i^2}(V_{ij}-X_{kj})^2-\log\sqrt{\frac{w_{ij}}{2\pi\sigma_i^2}}\Big)+\log u_{ik}\Big] \tag{4}$$

其中$\boldsymbol{u}_k=(u_{1k},u_{2k},\cdots,u_{ck})$是模糊隶属值组成的向量. 为了便于表达,下面给出了簇结构α-熵与聚类结构α-熵的定义.

定义1　设簇i的权向量为$\boldsymbol{w}_i=(w_{i1},w_{i2},\cdots,w_{ip})$,则簇结构α-熵由式(5)计算.

$$E(\boldsymbol{w}_i)=\frac{1}{2^{1-\alpha}-1}\left(\sum_{j=1}^{p}w_{ij}^{\alpha}-1\right) \tag{5}$$

其中$\alpha\geq 0,\alpha\neq 1$.

定义 2 聚类结构α-熵定义为所有簇结构α-熵的加权和,并由式(6)计算.

$$H(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_c)=\sum_{i=1}^{c}E(\boldsymbol{w}_i)=\sum_{i=1}^{c}\frac{h_i}{2^{1-\alpha}-1}\left(\sum_{j=1}^{p}w_{ij}^{\alpha}-1\right) \tag{6}$$

其中h_i为加权系数.

可以看到,当$\alpha\to 1$时,则簇结构α-熵即为模糊熵,聚类结构α-熵为聚类模糊熵.

将聚类结构α-熵引入式(4),并结合约束条件,从而获得了结构α-熵的加权高斯混合子空间聚类模型,如式(7)所示.

$$J(\boldsymbol{u}_k,\beta_i,\boldsymbol{V}_i,\sigma_i^2,w_{ij})=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\left[-\log\beta_i+\sum_{j=1}^{p}\left(\frac{w_{ij}}{2\sigma_i^2}(V_{ij}-x_{kj})^2-\log\sqrt{\frac{w_{ij}}{2\pi\sigma_i^2}}+\log u_{ik}\right]-\sum_{i=1}^{c}\frac{h_i}{2^{1-\alpha}-1}\left(\sum_{j=1}^{p}w_{ij}^{\alpha}-1\right)\right) \tag{7}$$

$$\text{s. t.}\ \left(\sum_{j=1}^{p}w_{ij}=1,0\leq w_{ij}\leq 1;\ \sum_{i=1}^{c}\beta_i^p=1,0\leq\beta_i\leq 1;\ \sum_{i=1}^{c}u_{ik}=1,0\leq u_{ik}\leq 1\right)$$

由式(7)可知,通过引入结构α-熵,并利用最小化KL散度以及最大化聚类结构α-熵,使得算法能够获取质量较高的簇,权向量的聚类结构α-熵值越大,则越表明有更多的维度用于发现聚类中的簇,从而得到更好的子空间聚类结构.

4 结构α-熵的加权高斯混合模型子空间聚类算法

针对式(7),构造拉格朗日函数,如式(8)所示.

$$L(\boldsymbol{u}_k,\beta_i,\boldsymbol{V}_i,\sigma_i^2,\boldsymbol{w}_i)=\frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\left[-\log\beta_i+\sum_{j=1}^{p}\left(\frac{w_{ij}}{2\sigma_i^2}(V_{ij}-x_{kj})^2-\log\sqrt{\frac{w_{ij}}{2\pi\sigma_i^2}}\right)+\log u_{ik}\right]-\sum_{i=1}^{c}\frac{h_i}{2^{1-\alpha}-1}\left(\sum_{j=1}^{p}w_{ij}^{\alpha}-1\right)+\sum_{i=1}^{c}\lambda_i\left(\sum_{j=1}^{p}w_{ij}-1\right)+\xi\left(\sum_{i=1}^{c}\beta_{ij}^{p}-1\right)+\sum_{k=1}^{n}\eta_k\left(\sum_{i=1}^{c}u_{ik}-1\right) \tag{8}$$

其中ξ为拉格朗日乘子,$\boldsymbol{\lambda}$和$\boldsymbol{\eta}$分别是由拉格朗日乘子$\lambda_1,\lambda_2,\cdots,\lambda_c$和$\eta_1,\eta_2,\cdots,\eta_n$构成的向量.

将式(8)分别对V_{ij}、σ_i和β_i求偏导,并令其等于0,从而得到式(9)~(11).

$$V_{ij}=\frac{\sum_{k=1}^{n}u_{ik}x_{kj}}{\sum_{k=1}^{n}u_{ik}} \tag{9}$$

$$\sigma_i^{\,2}=\frac{\sum_{k=1}^{n}u_{ik}\sum_{j=1}^{p}w_{ij}(V_{ij}-x_{kj})^2}{p\sum_{k=1}^{n}u_{ik}} \tag{10}$$

$$\beta_i=\left(\frac{1}{n}\sum_{k=1}^{n}u_{ik}\right)^{\frac{1}{p}} \tag{11}$$

对于隶属度u_{ik}和特征权值w_{ij}的求解,下面以定理形式给出.

定理 1 给定混合系数β_i与权值向量$\boldsymbol{w_i}=(w_{i1},w_{i2},\cdots,w_{ip})$,则式(8)取极值时满足的必要条件如式(12)所示.

$$u_{ik}=\frac{\beta_i\left(\prod_{j=1}^{p}\sqrt{\frac{w_{ij}}{2\pi\sigma_i^{\,2}}}\right)\exp\left(-\frac{1}{2\sigma_i^{\,2}}\sum_{j=1}^{p}w_{ij}(V_{ij}-x_{kj})^2\right)}{\sum_{t=1}^{c}\beta_t\left(\prod_{j=1}^{p}\sqrt{\frac{w_{ij}}{2\pi\sigma_t^{\,2}}}\right)\exp\left(-\frac{1}{2\sigma_t^{\,2}}\sum_{j=1}^{p}w_{tj}(V_{tj}-x_{kj})^2\right)} \tag{12}$$

定理 2 给定混合系数β_i以及隶属度u_{ik},则式(8)取极值时满足的必要条件如式(13)所示.

$$w_{ij}=\frac{\left(\frac{\alpha h_i}{2^{1-\alpha}-1}-Y_{ij}\right)^{\frac{1}{1-\alpha}}}{\sum_{j=1}^{p}\left(\frac{\alpha h_i}{2^{1-\alpha}-1}-Y_{ij}\right)^{\frac{1}{1-\alpha}}},\quad i=1,2,\cdots,c;\ j=1,2,\cdots,p \tag{13}$$

其中$Y_{ij}=\dfrac{\frac{1}{2n}\sum_{k=1}^{n}u_{ik}\left[\frac{(V_{ij}-x_{kj})^2}{\sigma_i^{\,2}}-\frac{1}{w_{ij}}\right]}{w_{ij}^{\alpha-1}}$,系数$h_i$可利用$h_i=\max_j(\text{abs}(Y_{ij}))\times\delta$确定[3],其中$\delta$为常数.

下面给出结构α-熵的加权高斯混合模型的子空间聚类算法SEWMM,如算法1.

算法1 结构α-熵的加权高斯混合模型的子空间聚类算法SEWMM

输入:数据集X,簇数c,参数α

输出:隶属矩阵$\boldsymbol{U}$,簇中心矩阵$\boldsymbol{V}$,权矩阵$\boldsymbol{W}$和方差σ_i,其中矩阵$\boldsymbol{U}$、$\boldsymbol{V}$与$\boldsymbol{W}$中的元素分别为u_{ik}、V_{ij}与$w_{ij}(i=1,2,\cdots,c;\ j=1,2,\cdots,p;\ k=1,2,\cdots,n)$

Step 1. 初始化$w_{ij}=1/p$,设置β_i和σ_i为常数,随机初始化中心矩阵V;

Step 2. 根据式(12)计算隶属度矩阵u_{ik};

Step 3. 根据式(10)计算方差σ_i;

Step 4. 根据式(13)计算权值w_{ij};

Step 5. 根据式(11)计算混合权值β_i;

Step 6. 根据式(9)计算中心V_{ij};

Step 7. 若目标函数达到最小值,则算法结束;否则,转Step 2.

可以看到,算法SEWMM包括两种类型的参数,一种是具有约束的参数,例如w_{ij}、β_i和u_{ik},另一种是不具有约束的参数,例如σ_i和$\boldsymbol{V}_{i\cdot}$.为了表明算法的收敛性,对于有约束参数,将式(8)关于w_{ij}、β_i和u_{ik}求二阶偏导数,分别得到式(14)~(16).

$$\frac{\partial^2 L}{\partial w_{ij}^2}=\frac{u_{ik}}{nw_{ij}^2}-\frac{\alpha(\alpha-1)}{2^{1-\alpha}-1}h_i w_{ij}^{\alpha-2} \tag{14}$$

$$\frac{\partial^2 L}{\partial \beta_i^2}=\frac{\sum_{k=1}^{n}u_{ik}}{n\beta_i^2}+\xi p(p-1)\beta_i^{p-2}>0 \tag{15}$$

$$\frac{\partial^2 L}{\partial u_{ik}^2}=\frac{1}{nu_{ik}}>0 \tag{16}$$

由$\alpha>0$与$\alpha\neq1$知,式(14)大于零,且式(15)与(16)也大于零.

对于无约束参数σ_i和$\boldsymbol{V}_i$,它们为极小值的证明可参考EM算法收敛性证明方法.

由以上结果可知,对于第t次迭代,则有

$$J(\boldsymbol{u}_k(t),\beta_i(t),\boldsymbol{V}_i(t),\sigma_i^2(t),w_{ij}(t))\leq$$
$$J(\boldsymbol{u}_k(t-1),\beta_i(t-1),\boldsymbol{V}_i(t-1),\sigma_i^2(t-1),w_{ij}(t-1))$$

此式表明函数$J(\boldsymbol{u}_k(t),\beta_i(t),\boldsymbol{V}_i(t),\sigma_i^2(t),w_{ij}(t))$关于变量$t$是一个非增函数,从而可知算法是收敛的.假设算法迭代T次收敛,由算法可知,其时间复杂性为$O(npTc)$.

5　实验研究

为了验证算法SEWMM的有效性,实验选取了UCI标准数据集与图像数据集,聚类评价指标分别为正确率(Accuracy,Acc)、兰德指数(Rand Index,RI)和标准互信息(Normalized Mutual Information,NMI),它们的取值范围均为[0,1];参数α的取值范围为[0.2,25],δ=10,实验结果为十次聚类的平均值.

5.1　UCI数据集

在UCI标准数据集[24]中,选取了Sonar、WBCD、SPECT Heart、Vehicle、Semeion、Water Treatment Plant(WTP)、Ionosphere、Iris、Mfeat_zer(MFeat的子数据集)和Liver数据集进行了实验,并与ESSC[25]、FWKM[15]、EWKM[18]、CKS-EWFC-F[26]、PFSCM[14]、kSDC[13]、MoG[23]、RGMM[22]和EWMM[3]算法进行了聚类性能比较,实验结果如表1所示,其中每个数据集所对应的三行数值分别为Acc、RI和NMI指标.可以看出,算法SEWMM在大部分数据集上获得了较好的聚类结果.例如,对于数据集Ionosphere、Semeion、Iris、WBCD、Vehicle与Liver,在选取的三个指标上,提出的算法均获得了较高值,而在Vehicle和WBCD数据集的Acc指标略低于MoG和EWKM算法,但其值基本相当.对于剩余的四个数据集,在选取的Acc、RI和NMI中的两个指标都超过了比较的算法,而对于另一指标,提出的算法与获得较好结果算法的性能也基本相当.同时,提出的算法在十个数据集的聚类性能优于EWMM.以上表明,通过引入结构α-熵惩罚项,提出的算法在Acc、NMI与RI上优于比较的算法.

5.2　图像数据集

本小节主要选取图像数据集进行实验和比较,包括BSDS500图像与较大规模图像数据集,实验中使用的图像均为彩色图像转换后的灰度图像.

5.2.1　BSDS500图像

针对BSDS500图像数据集[27],选取其中的16幅图像对提出的算法SEWMM进行了图像分割实验,比较算法分别为NCuts[28](记为NCuts1)、多尺度NCuts[29](记为NCuts2)、FLICM[30]、DSFCM-N[31]、FSC[16]、kSDC[13]、CKS-EWFC-F[26]、PFSCM[14]和EWMM[3].为了衡量算法的总体性能,在实验中,将不同算法分别在16幅图像获得的Acc、RI和NMI指标值进行了平均,实验结果如图1所示.可以看到,对于选取的所有图像,在聚类指标Acc和RI上,算法SEWMM均优于比较的算法,而在NMI指标上,算法SEWMM略低于NCust2;另外,我们也看到,在所有三个聚类指标上,提出的算法SEWMM的指标值都高于算法EWMM.

5.2.2　图像数据集的聚类

本小节选取CIFAR10[32]、CIFAR100[32]、MNIST[33]、Fashion-MNIST[34]和USPS[35]图像测试集对提出的算法进行了实验;同时,选取NMF[36]、SSC[20]、SSC-OMP[37]、LRR[21]、LSR[38]、LRSC[39]、EnSC[40]、CKS-EWFC-F[26]、PFSCM[14]、kSDC[13]、MoG[23]、RGMM[22]和EWMM[3]算法与提出的算法进行了实验比较.使用PCA方法对灰度图像进行特征提取,其中CIFAR-10、CIFAR-100、MNIST与Fashion-MNIST图像数据的特征数均置为500;对于USPS图像数据直接使用所有的灰度值,实验结果如表2所示,其中每个数据集所对应的三行数值分别为Acc、NMI和RI指标.

由表2可知,对于选取的图像数据集,算法SEWMM的聚类指标值优于EWMM算法;与其他聚类算法相比,总体上来说,算法SEWMM的聚类性能优于所比较的算法,例如,对于Fashion-MNIST数据集,算法SEWMM的Acc、NMI与RI指标值分别为0.5880、0.6139和0.8836,而EnSC算法的相应指标值分别为0.5777、0.6087和0.8935,可以看到,提出的算法在Acc和NMI两个指标均获得了较高值,而EnSC在RI指标上获得了较高值,但SEWMM与EnSC两种算法的RI指标值相差幅度不大.总之,在提出的算法中,通过引入结构α-熵,调整参数α的取值,提出的算法能够获得更好的聚类性能.

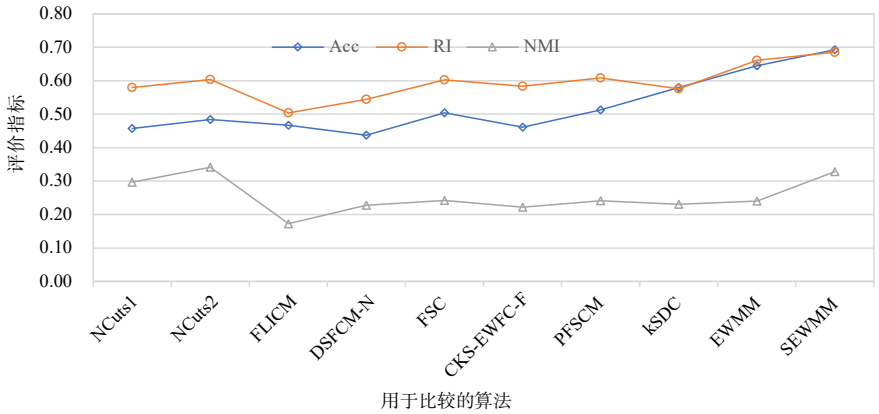

图1 图像分割实验结果

表1 不同算法在UCI数据集上的聚类性能比较

数据集	ESSC	FWKM	EWKM	CKS-EWFC-F	PFSCM	kSDC	MoG	RGMM	EWMM	SEWMM
Ionosphere	0.6685	0.7079	0.4803	0.7123	0.7094	0.7111	0.7164	0.7202	0.7094	0.7319
	0.6834	0.7199	0.5197	0.5889	0.5865	0.5880	0.6409	0.5959	0.5865	0.7636
	0.8498	0.4233	0.8197	0.1349	0.1299	0.1338	0.2507	0.1267	0.2606	0.8616
Liver	0.9679	0.9679	0.9679	0.9487	0.9423	0.9487	0.9551	0.9487	0.9688	0.9688
	0.9680	0.9680	0.9680	0.9021	0.8906	0.9021	0.9137	0.9488	0.9691	0.9688
	0.8266	0.8267	0.8267	0.7569	0.7363	0.7569	0.7786	0.7569	0.8803	0.8833
Mfeat_zer	0.1395	0.5340	0.5270	0.2416	0.3457	0.5458	0.5050	0.5220	0.5425	0.5560
	0.1924	0.5724	0.5793	0.6291	0.6425	0.7238	0.7893	0.7884	0.5951	0.6007
	0.7035	0.7792	0.7810	0.1199	0.3810	0.5029	0.5590	0.5282	0.7844	0.7894
WTP	0.3541	0.2349	0.2068	0.2019	0.1674	0.1803	0.2258	0.1973	0.3269	0.3496
	0.6301	0.6404	0.6392	0.6364	0.6315	0.6270	0.6420	0.6240	0.5924	0.7134
	0.1450	0.1640	0.1601	0.1271	0.0859	0.0849	0.1618	0.0959	0.3254	0.3291
Vehicle	0.3986	0.4188	0.4507	0.4301	0.4492	0.4508	0.4752	0.4468	0.4515	0.4508
	0.5543	0.6482	0.6378	0.6604	0.6506	0.6514	0.6701	0.6165	0.6691	0.6708
	0.1475	0.1680	0.1871	0.1881	0.1800	0.1851	0.1756	0.2068	0.1922	0.2423
Semeion	0.5874	0.6202	0.5545	0.3051	0.2751	0.5807	0.6246	0.6076	0.6284	0.6334
	0.8226	0.8945	0.8855	0.7171	0.6652	0.8840	0.8975	0.8985	0.8929	0.8959
	0.5543	0.5652	0.5374	0.3028	0.2845	0.5490	0.5212	0.5609	0.6100	0.5732
Sonar	0.5529	0.5409	0.5505	0.5553	0.5529	0.5538	0.5568	0.5481	0.5385	0.5702
	0.5110	0.5020	0.5043	0.5037	0.5032	0.5035	0.5261	0.5022	0.5006	0.5075
	0.0877	0.0071	0.0111	0.0097	0.0088	0.0094	0.0483	0.0081	0.0061	0.0124
Iris	0.8049	0.6551	0.7921	0.9033	0.8933	0.8867	0.9167	0.9101	0.9000	0.9200
	0.8695	0.9003	0.8667	0.8895	0.8797	0.8737	0.8366	0.9267	0.8848	0.9105
	0.7423	0.7882	0.7416	0.4536	0.7496	0.7419	0.8341	0.8334	0.7904	0.8505
WBCD	0.7116	0.6552	0.9571	0.8496	0.8541	0.8541	0.9121	0.8629	0.9371	0.9565
	0.5869	0.8365	0.8365	0.7508	0.7504	0.7504	0.8394	0.7630	0.8823	0.9156
	0.1225	0.5638	0.5944	0.4536	0.4672	0.4672	0.5621	0.4780	0.7207	0.7828
SPECT Heart	0.4382	0.4880	0.4491	0.3051	0.6030	0.5300	0.7191	0.6067	0.8432	0.8573
	0.5058	0.5067	0.5036	0.7171	0.5194	0.5087	0.5945	0.5210	0.7444	0.7533
	0.1097	0.0974	0.0952	0.3028	0.1029	0.0984	0.0718	0.0964	0.2494	0.2577

表2　不同算法对图像数据集的聚类性能比较

算法	CIFAR-100	CIFAR-10	MNIST	Fashion-MNIST	USPS
NMF	0.0926	0.2150	0.4865	0.4981	0.4252
	0.2341	0.1080	0.4551	0.5345	0.3810
	0.9791	0.8223	0.8693	0.8720	0.8577
SSC	0.0283	0.1220	0.3792	0.5118	0.2195
	0.2467	0.0489	0.4682	0.5764	0.2274
	0.9794	0.2189	0.8247	0.8751	0.7535
SSC-OMP	0.0129	0.1092	0.3863	0.4259	0.2328
	0.0392	0.0165	0.4740	0.5015	0.2021
	0.0792	0.1965	0.8262	0.8458	0.6847
LRR	0.0528	0.1012	0.4537	0.4674	0.5489
	0.1368	0.0210	0.4960	0.5717	0.4701
	0.8849	0.1307	0.7883	0.8194	0.8833
LSR	0.0861	0.2267	0.4929	0.4759	0.5315
	0.2069	0.1049	0.4968	0.5328	0.4707
	0.9775	0.8212	0.8682	0.8781	0.8792
LRSC	0.0794	0.1938	0.5336	0.5378	0.5420
	0.2202	0.0723	0.4964	0.5250	0.4830
	0.9787	0.8166	0.8813	0.8807	0.8814
EnSC	0.0823	0.2031	0.6122	0.5777	0.5178
	0.2139	0.0765	0.6606	0.6087	0.4977
	0.9787	0.8255	0.9048	0.8935	0.8714
CKS-EWFC-F	0.0322	0.1771	0.3485	0.2894	0.1002
	0.0897	0.0423	0.2560	0.2368	0.0066
	0.9185	0.8098	0.7772	0.6677	0.1002
PFSCM	0.0268	0.1496	0.2123	0.2854	0.1837
	0.0655	0.0445	0.1318	0.2636	0.1679
	0.6492	0.5103	0.5321	0.7531	0.5346
kSDC	0.0926	0.1941	0.2705	0.4853	0.4590
	0.2274	0.0899	0.4252	0.5151	0.4502
	0.9784	0.8225	0.5219	0.8733	0.8713
MoG	0.0906	0.2220	0.5134	0.5470	0.5210
	0.2266	0.0720	0.4907	0.5712	0.5433
	0.9792	0.8261	0.9443	0.8914	0.8750
RGMM	0.0853	0.2181	0.2658	0.4981	0.4528
	0.2224	0.1131	0.4447	0.5468	0.4906
	0.9779	0.8208	0.5262	0.8775	0.8722
EWMM	0.0895	0.2170	0.5152	0.5453	0.4401
	0.2255	0.0994	0.5062	0.5276	0.4524
	0.9785	0.8268	0.8754	0.8830	0.8647
SEWMM	0.0896	0.2250	0.5467	0.5880	0.4438
	0.2317	0.0961	0.5090	0.6139	0.4488
	0.9793	0.8273	0.8869	0.8836	0.8876

6　结论

本文在定义簇结构熵与聚类结构熵的基础上,将权向量的负结构α-熵引入到高斯混合模型中,提出了结构α-熵的加权高斯混合模型的子空间聚类算法SEWMM. 通过选取聚类正确率Acc、兰德指数RI与标准互信息NMI评价指标,在UCI数据集与图像数据集上,对提出的算法进行了实验研究. 实验结果表明,与已有的子空间聚类算法相比,算法SEWMM对数据的聚类效果更好. 同时,算法SEWMM的提出也为大数据背景下数据的聚类提供了一种新方法.

参考文献

[1] LI L, HANSMAN R J, PALACIOS R, et al. Anomaly detection via a Gaussian mixture model for flight operation and safety monitoring[J]. Transportation Research Part C, 2016, 64: 45-57.

[2] ZHAO Y, ABHISHEK K S, KWOK L T. Regularized Gaussian mixture model for high-dimensional clustering [J]. IEEE Transactions on Cybernetics, 2019, 49(10): 3677-3688.

[3] PENG L, ZHANG J Y. An entropy weighting mixture model for subspace clustering of high-dimensional data[J]. Pattern Recognition Letters, 2011, 32(8): 1154-1161.

[4] JYOTI M, ANGSHUL M, EMILIE C, et al. Deeply transformed subspace clustering[J]. Signal Processing, 2020, 174(107628): 1-6.

[5] LI H, ZHANG K, JIANG T. Minimum entropy clustering and applications to gene expression analysis[C]//Proceedings of the IEEE Computational Systems Bioinformatics Conference. Stanford: IEEE Computer Society, 2004: 142-151.

[6] NICHOLAS J T, SUYASH P A, GANG S, et al. Point set registration using Havrda-Charvat-Tsallis entropy measures [J]. IEEE Transactions on Medical Imaging, 2011, 30(2): 451-460.

[7] SUN R H, CHOL R J. Subspace Gaussian mixture based language modeling for large vocabulary continuous speech recognition[J]. Speech Communication, 2020, 117: 21-27.

[8] ZHAO L, SHANG Z, TAN J, et al. Adaptive parameter estimation of GMM and its application in clustering[J]. Future Generation Computer Systems, 2020, 106: 250-259.

[9] COSTEIRA J P, KANADE T. A multibody factorization method for independently moving objects[J]. International Journal of Computer Vision, 1998, 29(3): 159-179.

[10] WANG S, GUO W. Robust co-clustering via dual local learning and high-order matrix factorization[J]. Knowledge-Based System, 2017, 138: 176-187.

[11] VIDAL R, MA Y, SASTRY S. Generalized principal

component analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1945-1959.

[13] LI C N, SHAO Y H, GUO Y R, et al. Robust *k*-subspace discriminant clustering[J]. Applied Soft Computing Journal, 2019, 85(105858): 1-13.

[14] ARTHUR G, MARIE J L, CHRISTOPHE M. A proximal framework for fuzzy subspace clustering[J]. Fuzzy Sets and Systems, 2019, 366: 34-45.

[15] JING L, NG M K, XU J, et al. Subspace clustering of text documents with feature weighting *k*−means algorithm[C]// Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining. Hanoi: Springer, 2005: 802-812.

[16] GAN G, WU J. A convergence theorem for the fuzzy subspace clustering(FSC) algorithm[J]. Pattern Recognition, 2008, 41 (6) : 1939-1947.

[17] WANG J, WANG S T, CHUNG F, et al. Fuzzy partition based soft subspace clustering and its application in high dimensional data[J]. Information Science, 2013, 246: 133-154.

[18] JING L, NG M K, HUANG J Z. An entropy weighting *k*-means algorithm for subspace clustering of high dimensional sparse data[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8): 1-16.

[19] ZHOU J, CHEN L, PHILIPCHEN C L, et al. Fuzzy clustering with the entropy of attribute weights[J]. Neurocomputing, 2016, 198: 125-134.

[20] ELHAMIFAR E, VIDAL R. Sparse subspace clustering: algorithm, theory, and applications[J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 2013, 35(11): 2765-2781.

[21] LIU G, LIN Z, YAN S, et al. Robust recovery of subspace structures by low-rank representation[J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 2013, 35(1): 171-184.

[22] LU C, FENG J, LIN Z, et al. Subspace clustering by block diagonal representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 487-501.

[23] LI B, LU H, ZHANG Y, et al. Subspace clustering under complex noise[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(4): 930-940.

[24] BLAKE C, MERZ C J. UCI repository for machine learning databases[DB/OL]. http://www.ics.uci.edu/mlearn/ml-repository. html, 1998.

[25] DENG Z, KUPSZE C, CHUNG F L, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information[J]. Pattern Recognition, 2010, 43(3): 767-781.

[26] WANG J, DENG Z. Distance metric learning for soft subspace clustering in composite kernel space[J]. Pattern Recognition, 2016, 52: 113-134.

[27] JIA H, CHEUNG Y M. Subspace clustering of categorical and numerical data with an unknown number of clusters[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(8): 3308-3325.

[28] SHI J B, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[29] COUR T, BENEZIT F, SHI J. Spectral segmentation with multiscale graph decomposition[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005: 1124-1131.

[30] KRINIDIS S, CHATZIS V. A robust fuzzy local information c-means clustering algorithm[J]. IEEE Transactions on Image Processing, 2010, 19(5): 1328-1337.

[31] ZHANG Y, BAI X, FAN R, et al. Deviation-sparse fuzzy c-means with neighbor information constraint[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(1): 185-199.

[32] KRIZHEVSKY A. Learning Multiple Layers of Features from Tiny Images[R]. Torono: University of Toronto, 2009.

[33] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[34] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms[DB/OL]. https://github. com/zalandoresearch/fashion-mnist, 2017.

[35] ROWEIS S. USPS handwritten digits data[DB/OL]. https://cs.nyu.edu/~roweis/data.html, 2018.

[36] CAI D, HE X F, WANG X H, et al. Locality preserving nonnegative matrix factorization[C]//BRUIJN J D. Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2009: 1010-1015.

[37] YOU C, ROBINSON D P, VIDAL R. Scalable sparse subspace clustering by orthogonal matching pursuit[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3918-3927.

[38] LU C Y, MIN H, ZHAO Z Q, et al. Robust and efficient subspace segmentation via least squares regression[C]// ISHANI C. Proceedings of the European Conference on Computer Vision. Florence: Springer, 2012: 347-360.

[39] VIDAL R, FAVARO P. Low rank subspace clustering[J]. Pattern Recognition Letters , 2014, 43: 47-61.

[40] YOU C, LI C, ROBINSON D, et al. Oracle based active set algorithm for scalable elastic net subspace clustering [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3928-3937.

作者简介

李　凯　男,1963年出生于河北省保定市.现为河北大学教授,从事机器学习、模式识别与数据挖掘等方面研究工作.
E-mail: likai@hbu.edu.cn

张可心　女,1994年出生于河北省邯郸市.从事机器学习与数据挖掘等方面研究工作.
E-mail: 306432139@qq.com