ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.8, August 2010, pp.1878−1888
doi: 10.3724/SP.J.1001.2010.03607

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于混合距离学习的双指数模糊 C 均值算法*

王 骏[1,2,3], 王士同[2,3+]

[1](南京理工大学 计算机科学与技术学院,江苏 南京 210094)

[2](江南大学 信息工程学院,江苏 无锡 214122)

[3](南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

Double Indices FCM Algorithm Based on Hybrid Distance Metric Learning

WANG Jun[1,2,3], WANG Shi-Tong[2,3+]

[1](School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

[2](School of Information Technology, Jiangnan University, Wuxi 214122, China)

[3](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: E-mail: wxwangst@yahoo.com.cn

Wang J, Wang ST. Double indices FCM algorithm based on hybrid distance metric learning. *Journal of Software*, 2010,21(8):1878−1888. http://www.jos.org.cn/1000-9825/3607.htm

Abstract: To learn a good distance metric without any class label information, an algorithm named HDDI-FCM (double indices fuzzy C-means with hybrid distance) is proposed in this paper. In detail, the unknown distance metric is firstly represented as the linear combination of several known distance metrics. Then the algorithm is executed to perform the clustering task as well as learn the most suitable metric simultaneously. To guarantee the convergence of the algorithm, the Steffensen iteration is introduced into the process of updating cluster centers. The selection of parameter for the algorithm is also discussed. The experimental results on a collection of UCI (University of California, Irvine) datasets demonstrate the effectiveness of the proposed algorithm.

Key words: distance metric learning; clustering; fuzzy C-means algorithm; hybrid distance metric; Steffensen iteration method

摘 要: 提出了一种基于 DI-FCM(double indices fuzzy C-means)算法框架的无监督距离学习算法——基于混合距离学习的双指数模糊 C 均值算法 HDDI-FCM(double indices fuzzy C-means with hybrid distance).数据集未知距离度量被表示为若干已有距离的线性组合,然后执行HDDI-FCM,在对数据集进行有效聚类的同时进行距离学习.为了保证迭代算法收敛,引入了 Steffensen 迭代法来改进计算簇中心点的迭代公式.讨论了算法中参数的选择.基于 UCI(University of California,Irvine)数据集的实验结果表明该算法是有效的.

关键词: 距离学习;聚类;模糊 C 均值算法;混合距离;Steffensen 迭代法

中图法分类号: TP181 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant Nos.60773206, 60704047, 90820002 (国家自然科学基金)

Received 2008-07-28; Revised 2008-11-27; Accepted 2009-03-05

在聚类分析过程中,需要根据数据点之间的相似或相异程度,对数据点进行区分和分类.因此,为数据集选择合适的距离度量来评价数据点之间的相似或相异程度,对聚类分析的效果至关重要.在此过程中,选用不同的距离度量,其效果相去甚远.然而在实际研究中,由于研究人员缺乏对数据集的认识,所以很难为数据集选择合适的距离度量.在众多的距离度量中,欧氏距离最为常用.但是,欧氏距离对噪声比较敏感[1].此外,欧氏距离仅适用于特征空间中超球结构的数据集,对超立方体结构、超椭球结构的数据集效果不太理想.因此,在诸如生物工程、计算机视觉等领域,欧氏距离并不是一个很好的选择.

如何为数据集选择合适的距离度量,这一直是困扰学术界的一个难题.近年来,在机器学习领域已经提出了多种方法来学习数据集中未知的距离.根据是否有先验的训练样本提供,距离学习可分为有监督距离学习[2−6]和无监督距离学习[7,8]两类.前者借助于带标记的训练样本集,学习到数据集中的未知距离表示;而后者则没有任何关于数据集标记的先验信息.

对于无监督距离学习来说,其主要思想是在保持数据之间的局部或全局几何特性的前提下,学习隐藏在数据集中的低维流形.目前,大多数方法都是在保证几何特性的前提下把数据集向低维流形投影.本文提出了一种方法.在此方法中,适合于数据集的未知距离度量被表示为若干已有距离的线性组合,然后执行一种基于 DI-FCM(double indices fuzzy C-means)框架的算法.在实现距离学习的同时,也得到了良好的聚类结果.

本文第 1 节提出一种表示数据集距离度量的新方法——基于线性组合的混合距离表示方法,并给出用于更新中心点的 Steffensen 迭代公式.第 2 节回顾基于欧氏距离的双指数模糊 C 均值算法 DI-FCM.在此基础上,第 3 节提出基于混合距离的双指数模糊 C 均值算法 HDDI-FCM(double indices fuzzy C-means with hybrid distance).第 4 节给出距离的 3 种组合形式,通过实验数据说明此方法的有效性.第 5 节给出结论.

1 基于线性组合的混合距离表示新方法

距离空间是数学中一个重要的基本概念.距离空间(metric space)是一种拓扑空间,其上的拓扑由指定的距离函数决定.设 X 是一个非空集,X 被称为距离空间,是指在 X 上定义了一个二元实值函数 $d(\boldsymbol{x},\boldsymbol{y})$满足以下 3 个条件:

(i) 非负性:$d(\boldsymbol{x},\boldsymbol{y})>0,\forall \boldsymbol{x}\neq\boldsymbol{y},d(\boldsymbol{x},\boldsymbol{x})=0$.

(ii) 对称性:$d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$.

(iii) 三角不等式:$d(\boldsymbol{x},\boldsymbol{y})\leqslant d(\boldsymbol{x},\boldsymbol{z})+d(\boldsymbol{z},\boldsymbol{y}),\forall \boldsymbol{z}$.

这里,d 称为 X 上的一个距离,以 d 为距离的距离空间记作(X,d).

定理. 如果 $d_k(\boldsymbol{x},\boldsymbol{y})$是一个距离,则其线性组合 $D(\boldsymbol{x},\boldsymbol{y})=\sum_{k=1}^{n}\omega_k d_k(\boldsymbol{x},\boldsymbol{y})$ 也是一个距离.

证明:条件(i)、条件(ii)易证.以下证明 $D(\boldsymbol{x},\boldsymbol{y})$满足三角不等式(iii):

$$D(\boldsymbol{x},\boldsymbol{y})=\sum_{k=1}^{K}\omega_k d_k(\boldsymbol{x},\boldsymbol{y})\leqslant\sum_{k=1}^{K}\omega_k(d_k(\boldsymbol{x},\boldsymbol{z})+d_k(\boldsymbol{z},\boldsymbol{y}))=\sum_{k=1}^{K}\omega_k d_k(\boldsymbol{x},\boldsymbol{z})+\sum_{k=1}^{K}\omega_k d_k(\boldsymbol{z},\boldsymbol{y})=D(\boldsymbol{x},\boldsymbol{z})+D(\boldsymbol{z},\boldsymbol{y}),$$

即 $D(\boldsymbol{x},\boldsymbol{y})\leqslant D(\boldsymbol{x},\boldsymbol{z})+D(\boldsymbol{z},\boldsymbol{y})$.因此,$D(\boldsymbol{x},\boldsymbol{y})$是一个距离.证毕. □

本文通过如下线性组合来表示数据集中的未知距离度量:

$$D(\boldsymbol{x},\boldsymbol{y})=\sum_{k=1}^{K}\omega_k^{\,p}d_k(\boldsymbol{x},\boldsymbol{y}) \tag{1}$$

$$\text{s.t.}\ \sum_{k=1}^{K}\omega_k^{q}=1 \tag{2}$$

其中,$p>q>0$,$D(\boldsymbol{x},\boldsymbol{y})$表示数据点 x 到数据点 y 的距离,$d_k(\boldsymbol{x},\boldsymbol{y})$是其第 k 个距离分量.

假设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$ 为 s 维欧氏空间 R^s 中的数据集,$\boldsymbol{x}_j(j=1,2,\ldots,n)$为特征向量,$\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_c\}\subset R^s$,$\boldsymbol{v}_i$ $(i=1,2,\ldots,c)$为第 i 个簇的中心点,μ_{ij} 表示第 j 个样本点属于第 i 个簇的模糊程度,m 是模糊指标,c 为该数据集上簇的数量.通常使用最小平方误差法来估计每个簇的中心点 $\boldsymbol{v}_i$.对于 $\boldsymbol{v}_i$,使下式取得最小值:

$$J(\boldsymbol{X},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}{}^{m}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)\ .$$

对 $\boldsymbol{v}_i(i=1,2,\ldots,c)$求导,有

$$\frac{\partial J(\boldsymbol{X},\boldsymbol{V})}{\partial \mathbf{v}_i}=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{m}\frac{\partial D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)}{\partial \mathbf{v}_i}=0\ ,\ i=1,2,\ldots,c \tag{3}$$

如果 $D(\boldsymbol{x}_j,\boldsymbol{v}_i)$取欧氏距离,求解较为方便,可以直接写出中心点 $\boldsymbol{v}_i$的显式表达式.但是,本文采用多个距离的线性组合表示 $D(\boldsymbol{x}_j,\boldsymbol{v}_i)$,而且距离分量未知,因此,直接写出其显式表达式就很困难了.本文采用迭代法来求解.将公式(1)代入公式(3),得到:

$$\frac{\partial J(\boldsymbol{X},\boldsymbol{V})}{\partial \boldsymbol{v}_i}=2\sum_{j=1}^{n}\mu_{ij}^{m}D(\boldsymbol{x}_j,\boldsymbol{v}_i)\sum_{k=1}^{K}\omega_k^p\frac{\partial d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)}{\partial \boldsymbol{v}_i}=0\ ,\ i=1,2,\ldots,c \tag{4}$$

假设迭代过程产生关于中心点 $\boldsymbol{v}_i$的迭代序列 $\{\boldsymbol{v}_i^{(l)}\}_{l=1}^{\infty}$,$\varphi(\cdot)$为相应的迭代函数,则构造如下定点迭代过程:

$$\boldsymbol{v}_i^{(l+1)}=\varphi(\boldsymbol{v}_i^{(l)})\ ,\ i=1,2,\ldots,c,\ l=1,2,\ldots \tag{5}$$

距离分量 $d_k(\boldsymbol{x}_j,\boldsymbol{v}_i)$的选取是任意的,因此不能保证迭代过程收敛.为了使算法收敛,本文采用 Steffensen 迭代法[9]进行迭代,从而使目标函数收敛到其局部极小值.具体方法如下:

$$\boldsymbol{y}_i^{(l)}=\varphi(\boldsymbol{v}_i^{(l)}) \tag{6}$$

$$\boldsymbol{z}_i^{(l)}=\varphi(\boldsymbol{y}_i^{(l)}) \tag{7}$$

$$\boldsymbol{v}_i^{(l+1)}=\boldsymbol{v}_i^{(l)}-\frac{(\boldsymbol{y}_i^{(l)}-\boldsymbol{v}_i^{(l)})^2}{\boldsymbol{z}_i^{(l)}-2\boldsymbol{y}_i^{(l)}+\boldsymbol{v}_i^{(l)}}\ ,\ i=1,2,\ldots,c \tag{8}$$

2 基于欧氏距离的双指数模糊 C 均值算法

Zadeh 提出模糊集的概念之后,模糊聚类的研究成为学术界的一大热点,其中最为引人注目的就是模糊 C 均值算法 FCM(fuzzy C-means)[10,11].当前,研究的热点多集中于固定模糊指标 m 值的算法研究,对于约束条件的研究却是一个空白.在标准 FCM 中,模糊指标 m 取值范围为 $m>1$.本文在此基础上对约束条件的幂指数进行推广,向约束条件中引入幂指数 r,从而得到一种新算法,本文称其为双指数模糊 C 均值算法(DI-FCM).DI-FCM 的算法性能由模糊指标 m 和约束条件的幂指数 r 共同决定.其意义在于,在理论上有效地扩展了 m 的取值范围,将模糊指标 m 的取值由原先的 $m>1$ 扩展到 $m>r>0$.

令 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$为 s 维空间中的有限数据集,$\boldsymbol{x}_k=\{x_{k1},x_{k2},\ldots,x_{ks}\},k=1,2,\ldots,n$.令 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n\}\subset R^s$,其中,$\boldsymbol{v}_i$表示第 i 个簇的中心.$m>0$ 为模糊指标,$r>0$,对任意整数 c 有 $2\le c\le n$.定义$\|\boldsymbol{x}_k-\boldsymbol{v}_i\|=\sqrt{\sum_{h=1}^{s}(x_{kh}-v_{ih})^2}$.与 FCM 类似,DI-FCM($m$,$r$)的目标函数定义为

$$J_{m,r}(\boldsymbol{U},\boldsymbol{V})=\sum_{k=1}^{n}\sum_{i=1}^{c}\mu_{ik}^{m}\|\boldsymbol{x}_k-\boldsymbol{v}_i\|^2,\quad m>0 \tag{9}$$

其中,$\boldsymbol{U}=[\mu_{ik}]_{c\times n}$为模糊划分矩阵,其元素$\mu_{ik}$满足如下 3 个约束条件:

$$\mu_{ik}\in[0,1],\ k=1,2,\ldots,n,\ i=1,2,\ldots,c \tag{10}$$

$$\sum_{i=1}^{c}\mu_{ik}^{r}=1\ ,\ m>r>0,\ k=1,2,\ldots,n \tag{11}$$

$$0<\sum_{k=1}^{n}\mu_{ik}<n\ ,\ i=1,2,\ldots,c \tag{12}$$

使用 Lagrange 乘子法,易得 $J_{m,r}$在约束条件(10)~(12)下取局部极小值的必要条件如下:

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n}\mu_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{n}\mu_{ij}^{m}},\ m>0,\ i=1,2,\ldots,c \tag{13}$$

$$\mu_{ik}=\left(\frac{\|\boldsymbol{x}_k-\boldsymbol{v}_i\|^{-\frac{2}{m/r-1}}}{\sum_{i=1}^{c}\|(\boldsymbol{x}_k-\boldsymbol{v}_i)\|^{-\frac{2}{m/r-1}}}\right)^{\frac{1}{r}},\ m>r>0,\ i=1,2,\ldots,c,\ k=1,2,\ldots,n \tag{14}$$

显然,当 r=1 时,DI-FCM 算法退化为 FCM 算法;当 $r\neq1$ 时,根据信息论的相关概念,数据集 X 中第 k 个样本的 r 阶β熵为

$$H_{\beta_r}(x_k)=\frac{1}{1-2^{r-1}}\left(1-\sum_{i=1}^{c}\mu_{ik}^{r}\right),\ r>0\ 且\ r\neq1 \tag{15}$$

显然,当 $r\neq1$ 且满足约束条件(11)时,$H_{\beta_r}(\boldsymbol{x}_k)$ 值为 0,这意味着数据集 $\boldsymbol{X}$ 中样本点 $\boldsymbol{x}_k$ 从属于各个簇的不确定性最小.故 DI-FCM(m,r)算法的实质就是数据集 $\boldsymbol{X}$ 中关于各个样本点 $\boldsymbol{x}_k$ 的模糊隶属度μ_{ik}的 r 阶β熵为 0 值的情况下,求解目标函数 $J_{m,r}$ 的最小值.

根据文献[10],为了保证算法收敛,目标函数(9)应该是 $\boldsymbol{U}$ 上的凸函数,满足 $m>r>0$.根据已有的研究,通常,m/r 在[1.5,2.5]之间时可以得到较好的聚类效果.

3 基于混合距离学习的双指数模糊 C 均值算法

3.1 HDDI-FCM算法及推导

对于聚类算法而言,建立合适的准则函数对算法的效果至关重要.从直观上来说,我们总希望类内距离应尽可能地小,而类间距离尽可能地大.在此基础上加以引申,即各个数据点到各自所属簇的中心的距离之和应尽可能地小.结合双指数模糊 C 均值算法和本文提出的基于线性组合的混合距离,本文定义准则函数及约束条件如下:

$$J(\boldsymbol{U},\boldsymbol{V})=\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^{m}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i) \tag{16}$$

其中:

$$D(\boldsymbol{x}_j,\boldsymbol{v}_i)=\sum_{k=1}^{K}\omega_k^{p}d_k(\boldsymbol{x}_j,\boldsymbol{v}_i) \tag{17}$$

为混合距离;$\boldsymbol{U}=[\mu_{ij}]_{c\times n}$ 为模糊划分矩阵,其元素μ_{ij}满足:

$$\sum_{i=1}^{c}\mu_{ij}^{r}=1 \tag{18}$$

距离分量 d_k 权重ω_k满足:

$$\sum_{k=1}^{K}\omega_k^{q}=1 \tag{19}$$

运用 Lagrange 乘子法,构造无约束条件的准则函数如下:

$$E=\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^{m}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)+\sum_{j=1}^{n}\gamma_j\left(\sum_{i=1}^{c}\mu_{ij}^{r}-1\right)+\lambda\left(\sum_{k=1}^{K}\omega_k^{q}-1\right) \tag{20}$$

上式取极小值的必要条件为

$$\frac{\partial E}{\partial u_{ij}}=m\mu_{ij}^{m-1}D^2(\boldsymbol{x}_j,\boldsymbol{v}_i)+\gamma_j r\mu_{ij}^{r-1}=0,\quad i=1,2,\ldots,c,\ j=1,2,\ldots,n \tag{21}$$