DOI:10.11992/tis.201504027
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.tp.20151111.1633.006.html

基于最大间隔理论的组合距离学习算法

郭瑛洁,王士同,许小龙
(江南大学 数字媒体学院,江苏 无锡 214000)

摘 要:从已知数据集中学习距离度量在许多机器学习应用中都起着重要作用。传统的距离学习方法通常假定目标距离函数为马氏距离的形式,这使得学习出的距离度量在应用上具有局限性。提出了一种新的距离学习方法,将目标距离函数表示为若干候选距离的线性组合,依据最大间隔理论利用数据集的边信息学习得到组合距离中各距离分量的权值,从而得到新的距离度量。通过该距离度量在模糊 C 均值聚类算法中的表现来对其进行评价。在 UCI 数据集上,与其他已有的距离学习算法的对比实验结果证明了该文算法的有效性。

关键词:距离学习;组合距离;最大间隔;FCM;模糊聚类;聚类算法;距离;学习算法
中图分类号:TP181 **文献标志码**:A **文章编号**:1673-4785(2015)06-0843-08

中文引用格式:郭瑛洁,王士同,许小龙. 基于最大间隔理论的组合距离学习算法[J]. 智能系统学报, 2015, 10(6): 843-850.
英文引用格式:**GUO Yingjie, WANG Shitong, XU Xiaolong. Learning a linear combination of distances based on the maximum-margin theory[J]. CAAI Transactions on Intelligent Systems**, 2015, 10(6): 843-850.

Learning a linear combination of distances based on the maximum-margin theory

GUO Yingjie, WANG Shitong, XU Xiaolong
(School of Digital Media, Jiangnan University, Wuxi 214000, China)

Abstract: Learning a distance metric from given training samples is a crucial aspect of many machine learning tasks. Conventional distance metric learning approaches often assume the target distance function to be represented in the form of Mahalanobis distance, and the metric has limitations for this application. This paper proposes a new metric learning approach in which the target distance function is represented as a linear combination of several candidate distance metrics. This method obtains a new distance metric by learning weights from side information according to the maximum-margin theory. The new distance function is applied to fuzzy C-means clustering for evaluation. The experiments were performed using UCI data, and a comparison of the results with those of other approaches reveals the advantages of the proposed technique.

Keywords: metric learning; hybrid distance metric; maximum-margin theory; fuzzy C-means; fuzzy clustering; clustering algorithm; metric; learning algorithm

如何表示 2 点之间的距离是模式识别中的基础问题。一个好的距离度量能够根据数据的结构与分布适用于不同的应用。欧氏距离是众多数据挖掘应用中使用最多的距离度量,但是欧氏距离仅适用于特征空间中超球结构的数据集,对于超立方体结构、超椭球结构的数据集效果不太理想[1]。除了欧氏距离,余弦距离是另一个应用广泛的距离度量。尽管余弦距离在文本检索中有优秀的表现,但是其预先假设了数据集每一维度都是等权重的[2],这一特性显然限制了余弦距离的应用范围。因此,欧式距离和余弦距离在实际应用中都不是最理想的选择。

从训练样本中学习出合适的距离度量是近年来

收稿日期:2015-04-19. 网络出版日期:2015-11-11.
基金项目:国家自然科学基金资助项目(61272210);江苏省自然科学基金资助项目(BK2011417, BK2011003).
通信作者:郭瑛洁. E-mail:ying_dm@163.com.

的研究热点,它对于提高聚类和分类效果有着重要的影响。一般的距离学习方法都是首先假定一个距离函数模型并进行求解,其中大部分的距离函数假定在马氏距离定义的框架之下,即对于 2 点 x、y,使用距离公式 $d(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})$,其中 $\boldsymbol{A}$ 为所要学习的距离矩阵。比如文献[3]中通过使相似样本之间距离减小学习了一个全局距离度量;区分成分分析(DCA)[4]通过最小化相似样本之间距离的同时最大化不相似样本之间的距离来学习距离矩阵;近邻成分分析(NCA)[5]通过最优化最近邻分类器的精度去学习马氏距离度量;最大边界近邻分类方法(LMNN)[6]在 NCA 的框架下拓展了最大边界的目标,但是学习的目标仍然是得到一个马氏距离。当马氏距离中的矩阵 $\boldsymbol{A}$ 取单位矩阵 $\boldsymbol{I}$ 时,则马氏距离表示欧氏距离。因此,本质上来说,以马氏距离为目标学习得到的新距离是欧式距离的线性变换,其无法准确地度量所有样本之间的距离。

有别于传统的距离学习方法,本文提出的距离学习方法并没有将学习目标单纯设定为马氏距离,而是学习由若干候选距离线性组合而成的新距离。基于最大间隔理论建立目标函数,利用数据的边信息通过对目标函数进行优化从而得到组合距离中的权重进而得到新的距离度量。对于候选距离的选择也不仅仅局限于马氏距离,本文选择了其他形式的距离度量进行组合,以扩大距离度量的适用范围。

1 组合距离表示

为了更好地表示数据点之间的距离,在距离函数中引入权重来强化有积极作用的部分,削减冗余的部分,已经成为一种常用的方法。在之前的方法中,研究者往往使用特征加权距离的方法[7]来改进聚类算法,特征加权距离的计算表达式为

$$w_{\mathrm{dist}}(\boldsymbol{x},\boldsymbol{y})=\sqrt{\sum_{h=1}^{d}\omega_h^{\alpha}(\boldsymbol{x}_h-\boldsymbol{y}_h)^2},\alpha>1$$

式中:$\boldsymbol{x}=[\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_d]^{\mathrm{T}}$,$\boldsymbol{y}=[y_1\ y_2\ \cdots\ y_d]^{\mathrm{T}}$ 为特征空间 $\boldsymbol{R}^d$ 中的任意 2 点,ω_h 为特征权重且满足 $\sum_{h=1}^{d}\omega_h=1$。

受特征加权距离的启发,引入权值将距离函数改写为若干候选距离的线性组合,将特征加权改为距离加权,从而强化对某一数据集有更好度量效果的距离分量。

本文通过以下线性组合来表示数据集中的距离度量:

$$D(\boldsymbol{x}_a,\boldsymbol{x}_b)=\sum_{i=1}^{p}\omega_i d_i(\boldsymbol{x}_a,\boldsymbol{x}_b)\qquad(1)$$

$$\text{s.t.}\sum_{i=1}^{p}\omega_i=1,$$

$$0\leqslant\omega_i\leqslant 1,\quad i=1,2,\cdots,p$$

式中:$D(\boldsymbol{x}_a,\boldsymbol{x}_b)$ 表示数据点 $\boldsymbol{x}_a$ 到数据点 $\boldsymbol{x}_b$ 之间的距离,它由 p 个距离分量组成,$d_i(\boldsymbol{x}_a,\boldsymbol{x}_b)$ 是其第 i 个距离分量,ω_i 是第 i 个距离分量所对应的权值。ω_i 需要满足各个分量权值均为正且和为 1 的条件。

在距离分量的选择上,除了经典的欧式距离之外,本文选择了若干含有数据维度方差的距离分量(如:$d(\boldsymbol{x}_a,\boldsymbol{x}_b)=(\boldsymbol{x}_a-\boldsymbol{x}_b)^{\mathrm{T}}\dfrac{\boldsymbol{I}}{\beta\sigma^2}(\boldsymbol{x}_a-\boldsymbol{x}_b)$,其中 β 为常数,$\boldsymbol{I}$ 为单位矩阵,σ 为数据点之间的标准差)以保留数据各特征分量上的特征。但是这些距离均为马氏距离定义框架下的距离度量,对其进行线性组合后,得到的距离函数仍为马氏距离形式。因此,根据 Wu 等提出的新距离[8],本文给出了若干形如 $d(\boldsymbol{x}_a,\boldsymbol{x}_b)=1-\exp(-\beta\parallel\boldsymbol{x}_a-\boldsymbol{x}_b\parallel^2)$ 的距离分量进行组合,其中 β 为常数,这些距离均为非线性的距离度量,通过组合可以形成非线性的距离函数以克服马氏距离的缺点。

2 基于最大间隔理论的距离学习

2.1 距离学习方法

本文所提出的距离学习算法将利用数据集的边信息进行学习,而边信息通常以成对约束的形式表现。因此,本文以成对约束的集合作为训练集并表示为 $D=\{(\boldsymbol{x}_a^k,\boldsymbol{x}_b^k,y_k),k=1,2,\cdots,n\}$,其中 n 为成对约束的对数。D 中每一对成对约束 $(\boldsymbol{x}_a^k,\boldsymbol{x}_b^k,y_k)$ 都是一个包含三个元素的元组,其中 $\boldsymbol{x}_a^k$ 和 $\boldsymbol{x}_b^k$ 是被表示为 d 维向量的样本点,y_k 是表示样本点 $\boldsymbol{x}_a^k$ 和 $\boldsymbol{x}_b^k$ 之间关系的类标。当 $\boldsymbol{x}_a^k$ 和 $\boldsymbol{x}_b^k$ 为同一类的样本点时,y_k 为正(如:$y_k=+1$);反之,y_k 为负(如:$y_k=-1$)。使用 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$ 来表示 D 中出现的所有的训练样本点,其中 N 表示样本点的个数。

统计学习中常用的经验风险最小化并不能保证良好的泛化性能,因此间隔理论[9]就伴随着过拟合的问题研究被提出,并逐渐成为机器学习领域中的一个重要评价标准。本文依据最大间隔理论并受 L2-SVM 方法[10]和文献[2]的启发,构建目标函数如式(2):

$$\min J=\frac{1}{2}\sum_{i=1}^{p}\omega_i^2+C\sum_{k=1}^{n}\xi_k^2+\beta^2-\theta\rho$$

$$\text{s.t.}\,y_k\Big(\sum_{i=1}^{p}\omega_i d_i(\boldsymbol{x}_a^k,\boldsymbol{x}_b^k)-\beta\Big)\geqslant\rho-\xi_k\qquad(2)$$

$$\sum_{i=1}^{p}\omega_i=1,0\leqslant\omega_i\leqslant 1,i=1,2,\cdots,p$$

式中:$d_i(\boldsymbol{x}_a^k,\boldsymbol{x}_b^k)$ 表示第 k 对成对约束的第 i 个距离分量,为了便于表示,本文在后续的介绍中将使用符号 $d_{i,k}$ 来代替 $d_i(\boldsymbol{x}_a^k,\boldsymbol{x}_b^k)$。此外,$y_k$ 为该约束对的类标,C 为惩罚因子,C 值大时对训练错误的惩罚增大,θ 为已知参数,β 为阈值,最大间隔为 $\frac{\rho}{\|\boldsymbol{\omega}\|}$。在优化的过程中最大化 ρ,最小化 $\|\boldsymbol{\omega}\|^2$,并使训练误差 ξ_k 最小化,其中 $\xi_k \geqslant 0$。

本文的目标是通过优化该目标函数求得距离分量的权值 ω_i,下面将具体介绍求解 ω_i 的方法。为了求解上述优化问题,将它作为原始最优化问题,应用拉格朗日对偶性,通过求解对偶问题得到原始问题的最优解。接下来将介绍具体求解过程。

首先,构建拉格朗日函数如式(3):

$$L=\frac{1}{2}\sum_{i=1}^{p}\omega_i^2+C\sum_{k=1}^{n}\xi_k^2+\beta^2-\theta\rho+\sum_{k=1}^{n}\alpha_k\Big(\rho-\xi_k-y_k\Big(\sum_{i=1}^{p}\omega_i d_{i,k}-\beta\Big)\Big)+\lambda\Big(1-\sum_{i=1}^{p}\omega_i\Big) \tag{3}$$

式中:$\boldsymbol{\alpha}=[\alpha_1\ \ \alpha_2\ \ \cdots\ \ \alpha_n]^{\mathrm{T}}$、$\lambda$ 均为拉格朗日乘子。

如果此时考虑式(2)中 $\omega_i \geqslant 0$ 的约束条件,并将该条件加入拉格朗日函数,则得到

$$L'=L-\sum_{i=1}^{p}\varphi_i\omega_i$$

将该拉格朗日函数分别对 ω_i、β、ρ、ξ_k、λ 求偏导,并令其等于 0 得到

$$\begin{cases}\dfrac{\partial L'}{\partial\omega_i}=\omega_i-\sum\limits_{k=1}^{n}\alpha_k y_k d_{i,k}-\lambda-\varphi_i=0\\ \dfrac{\partial L'}{\partial\beta}=2\beta-\sum\limits_{k=1}^{n}\alpha_k y_k=0\\ \dfrac{\partial L'}{\partial\rho}=-\theta+\sum\limits_{k=1}^{n}\alpha_k=0\\ \dfrac{\partial L'}{\partial\xi_k}=2C\xi_k-\alpha_k=0\\ \dfrac{\partial L'}{\partial\lambda}=1-\sum\limits_{i=1}^{p}\omega_i=0\end{cases} \tag{4}$$

显然由方程组(4)无法求得 ω_i,因此本文先暂时不考虑 $\omega_i \geqslant 0$ 的约束条件,使用式(3)进行后续的求解。

根据拉格朗日对偶性,原始问题的对偶问题是极大极小问题:

$$\max_{\alpha}\min_{\boldsymbol{\omega},\beta,\boldsymbol{\xi}}L(\boldsymbol{\omega},\beta,\boldsymbol{\xi},\boldsymbol{\alpha},\lambda)$$

所以,需要先求 $L(\boldsymbol{\omega},\beta,\boldsymbol{\xi},\boldsymbol{\alpha},\lambda)$ 对于 $\boldsymbol{\omega}$、λ 的极小值。

将拉格朗日函数式(5)分别对 ω_i、β、ρ、ξ_k、λ 求偏导,并令其等于 0 得到

$$\begin{cases}\dfrac{\partial L}{\partial\omega_i}=\omega_i-\sum\limits_{k=1}^{n}\alpha_k y_k d_{i,k}-\lambda=0\\ \dfrac{\partial L}{\partial\beta}=2\beta-\sum\limits_{k=1}^{n}\alpha_k y_k=0\\ \dfrac{\partial L}{\partial\rho}=-\theta+\sum\limits_{k=1}^{n}\alpha_k=0\\ \dfrac{\partial L}{\partial\xi_k}=2C\xi_k-\alpha_k=0\\ \dfrac{\partial L}{\partial\lambda}=1-\sum\limits_{i=1}^{p}\omega_i=0\end{cases} \tag{5}$$

进而得到

$$\omega_i=\sum_{k=1}^{n}\alpha_k y_k d_{i,k}+\lambda \tag{6}$$

$$\beta=\frac{1}{2}\sum_{k=1}^{n}\alpha_k y_k \tag{7}$$

$$\theta=\sum_{k=1}^{n}\alpha_k \tag{8}$$

$$\xi_k=\frac{\alpha_k}{2C} \tag{9}$$

$$1-\sum_{i=1}^{p}\omega_i=0 \tag{10}$$

将式(6)代入式(10)得到

$$\lambda=\frac{1}{p}\Big(1-\sum_{k=1}^{n}\alpha_k y_k d_{i,k}\Big) \tag{11}$$

进而,将式(11)代入式(6)得到

$$\omega_i=\frac{1}{p}-\frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k}+\sum_{k=1}^{n}\alpha_k y_k d_{i,k} \tag{12}$$

将式(7)~(10)、(12)代入拉格朗日函数(3)中,即得

$$L=\frac{1}{2p}-\frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k}+\frac{1}{2p}\Big(\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k}\Big)^2-\frac{1}{2}\sum_{i=1}^{p}\Big(\sum_{q=1}^{n}\alpha_q y_q d_{i,q}\sum_{r=1}^{n}\alpha_r y_r d_{i,r}\Big)+\sum_{q=1}^{n}\alpha_q y_q\sum_{r=1}^{n}\alpha_r y_r$$

即

$$\min_{\boldsymbol{\omega},\beta,\boldsymbol{\xi}}L(\boldsymbol{\omega},\beta,\boldsymbol{\xi},\boldsymbol{\alpha},\lambda)=\frac{1}{2p}-\frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k}+\frac{1}{2p}\Big(\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k}\Big)^2-\frac{1}{2}\sum_{i=1}^{p}\Big(\sum_{q=1}^{n}\alpha_q y_q d_{i,q}\sum_{r=1}^{n}\alpha_r y_r d_{i,r}\Big)+$$

$$\sum_{q=1}^{n}\alpha_q y_q \sum_{r=1}^{n}\alpha_r y_r$$

求 $\min_{\omega,\beta,\xi} L(\boldsymbol{\omega},\beta,\boldsymbol{\xi},\boldsymbol{\alpha},\lambda)$ 对 $\boldsymbol{\alpha}$ 的极大,即得对偶问题:

$$\max_{\alpha} \frac{1}{2p} - \frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k} +$$

$$\frac{1}{2p}\Big(\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k}\Big)^2 -$$

$$\frac{1}{2}\sum_{i=1}^{p}\Big(\sum_{q=1}^{n}\alpha_q y_q d_{i,q}\sum_{r=1}^{n}\alpha_r y_r d_{i,r}\Big) + \sum_{q=1}^{n}\alpha_q y_q \sum_{r=1}^{n}\alpha_r y_r$$

$$\text{s.t.}\sum_{k=1}^{n}\alpha_k y_k = 0$$

$$\sum_{k=1}^{n}\alpha_k = \theta$$

$$\alpha_k \geqslant 0, k = 1,2,\cdots,n \tag{13}$$

将式(13)的目标函数由求极大转换为求极小,就得到下面与之等价的对偶最优化问题:

$$\min_{\alpha} -\frac{1}{2p} + \frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k} -$$

$$\frac{1}{2p}\Big(\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k y_k d_{i,k}\Big)^2 +$$

$$\frac{1}{2}\sum_{i=1}^{p}\Big(\sum_{q=1}^{n}\alpha_q y_q d_{i,q}\sum_{r=1}^{n}\alpha_r y_r d_{i,r}\Big) - \sum_{q=1}^{n}\alpha_q y_q \sum_{r=1}^{n}\alpha_r y_r$$

$$\text{s.t.}\sum_{k=1}^{n}\alpha_k y_k = 0$$

$$\sum_{k=1}^{n}\alpha_k = \theta$$

$$\alpha_k \geqslant 0, k = 1,2,\cdots,n$$

如此,可以通过二次规划的求解方法得到最优解 $\boldsymbol{\alpha}^* = [\alpha_1^* \quad \alpha_2^* \quad \cdots \quad \alpha_n^*]^{\mathrm{T}}$,进而代入式(12)得到 $\boldsymbol{\omega}_i$ 的最优解:

$$\omega_i^* = \frac{1}{p} - \frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n}\alpha_k^* y_k d_{i,k} + \sum_{k=1}^{n}\alpha_k^* y_k d_{i,k}$$

可以明显地观察到,即使成功的优化得到最优解,也不能保证 $\boldsymbol{\omega}_i$ 完全满足式(2)中 $\boldsymbol{\omega}_i \geqslant 0$ 的约束条件,受 PFC 算法[11]的启发,在之前的基础上对 $\boldsymbol{\omega}_i$ 做如下修改:

$$\boldsymbol{\omega}_i = \begin{cases} 0, i \in p^- \\ \dfrac{1}{|p^+|} - \dfrac{1}{|p^+|}\displaystyle\sum_{j\in p^+}\sum_{k=1}^{n}\alpha_k^* y_k d_{j,k} + \sum_{k=1}^{n}\alpha_k^* y_k d_{i,k}, i \in p^+ \end{cases} \tag{14}$$

式中:

$$p^- = \{i : \boldsymbol{\omega}_i = 0\}$$

$$p^+ = \{i : \boldsymbol{\omega}_i > 0\}$$

p^+ 表示所有使得 $\boldsymbol{\omega}_i$ 取正值的 i 的集合,相对应的 p^- 表示无法使 $\boldsymbol{\omega}_i$ 取正值的 i 的集合,2 个集合 p^+ 和 p^- 的大小分别使用 $|p^+|$ 和 $|p^-|$ 来表示。

至此完成求解距离分量权值 $\boldsymbol{\omega}_i$ 的目标,求解集合 p^+ 和 p^- 的算法描述将在下一小节给出。

最大几何间隔 $\dfrac{\rho}{\|\boldsymbol{\omega}\|}$ 中的 ρ 亦可在求解权值 $\boldsymbol{\omega}_i$ 的过程中求得。与原始 SVM 类似,目标函数(3)的分离超平面为 $y_k(\sum_{i=1}^{p}\boldsymbol{\omega}_i d_{i,k} - \beta) = \rho$,即可得最大函数间隔 $\rho = y_k(\sum_{i=1}^{p}\boldsymbol{\omega}_i d_{i,k} - \beta)$。在求得最优解后由式(9)可得 β 的最优解 $\beta^* = \dfrac{1}{2}\sum_{k=1}^{n}\alpha_k^* y_k$ 进而可得最大函数间隔 $\rho = y_k(\sum_{i=1}^{p}\boldsymbol{\omega}_i{}^* d_{i,k} - \beta^*)$。

2.2 算法描述

本节将给出求解距离分量权值 $\boldsymbol{\omega}_i$ 的具体算法步骤。

为了便于表示,将式(14)简化:

$$\boldsymbol{\omega}_i = \begin{cases} 0, i \in p^- \\ \dfrac{1}{|p^+|} + CV_i, i \in p^+ \end{cases} \tag{15}$$

式中:

$$V_i = -\frac{1}{|p^+|}\sum_{j\in p^+}\sum_{k=1}^{n}\alpha_k^* y_k d_{j,k} + \sum_{k=1}^{n}\alpha_k^* y_k d_{i,k} \tag{16}$$

由式(16)观察可得,若想满足 $\boldsymbol{\omega}_i$ 为正值,则 V_i 需要足够大,且当 V_i 越大,$\boldsymbol{\omega}_i$ 为正值的几率就越大。因此,求解集合 p^+ 和 p^- 的算法总结如下。

算法 1 求解集合 p^+ 和 p^-。

1)初始化:$p_0^+ = \varnothing, p_0^- = \{1,2,\cdots,p\}, h = 0$;

2)$h = h + 1, p_h^+ = p_{h-1}^+ + \{i\}, p_h^- = p_{h-1}^- - \{i\}$,其中,$i = \arg\max_{i\in p_{h-1}^-}\{V_i\}$;

3)通过式(14)计算 $\boldsymbol{\omega}_g$ 并判断其是否大于 0。其中,$g = \arg\min_{i\in p_h^+}\{V_i\}$。如果 $\boldsymbol{\omega}_g > 0$ 则回到 2),否则设置 $p^+ = p_{h-1}^+, p^- = p_{h-1}^-$ 并终止。

下面将介绍学习距离函数中权值 $\boldsymbol{\omega}_i$ 的算法,其中 $\boldsymbol{\omega}_i$ 将采用如下方法初始化:在式(1)中 $\boldsymbol{\omega}_i$ 的约束条件下,令 $\boldsymbol{\omega}_1^{(0)} = \boldsymbol{\omega}_2^{(0)} = \cdots = \boldsymbol{\omega}_p^{(0)}$,因此有 $\boldsymbol{\omega}_i^{(0)} = 1/p$。

算法 2 学习距离函数。

输入:

1)数据矩阵:$\boldsymbol{X} \in \mathbf{R}^{d\times N}$;

2)成对约束:$(\boldsymbol{x}_a^k, \boldsymbol{x}_a^k, y_k)$ 其中,$y_k = \{+1, -1\}$;

3)参数:C、θ;

输出:距离权值:$\boldsymbol{\omega}$;

方法:

1)初始化:$\boldsymbol{\omega} = \boldsymbol{\omega}^{(0)}$;

2)计算距离矩阵:$\boldsymbol{D}(i,k)$;

3)计算二次规划参数 H 和 f:

$$H = \sum_{i=1}^{p}\left(\sum_{q=1}^{n}\sum_{r=1}^{n} d_{i,q}d_{i,r}y_q y_r\right) - \frac{1}{d}\left(\sum_{i=1}^{p}\sum_{k=1}^{n} y_k d_{i,k}\right)^2 + y_k^2$$

$$f = \frac{1}{p}\sum_{i=1}^{p}\sum_{k=1}^{n} y_k d_{i,k}$$

4)利用二次规划优化算法求解得到最优解:$\boldsymbol{\alpha}^* = [\alpha_1^* \quad \alpha_2^* \quad \cdots \quad \alpha_n^*]^{\mathrm{T}}$;

5)计算集合 p^+ 和 p^-:算法 1;

6)利用式(15)计算 $\boldsymbol{\omega}$。

3　实验

为了与传统 FCM 算法之间有可比性,本文将简单的以学习得到的距离函数替换传统 FCM 算法中的欧式距离。根据传统 FCM 算法的实现方法,本文将通过以下步骤实现聚类:

1)初始化隶属度矩阵 $\boldsymbol{U}$,使得 $\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,2,\cdots,n, u_{ij} \in [0,1]$。

2)计算聚类中心:$\boldsymbol{c}_i = \dfrac{\sum_{j=1}^{N} u_{ij}^m \boldsymbol{x}_j}{\sum_{j=1}^{N} u_{ij}^m}$。

3)计算价值函数:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^m d_{ij}^2$$

当其相对于上次价值函数值的改变量小于某个阈值时,算法停止。

4)更新隶属度矩阵:

$$\boldsymbol{u}_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\dfrac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$$

其中对于样本点 $\boldsymbol{x}_j$,它与聚类中心 $\boldsymbol{c}_i$ 之间的距离使用如下公式计算:

$$d_{ij} = \sum_{r=1}^{p} \omega_r d_r(\boldsymbol{x}_j, \boldsymbol{c}_i)$$

将上述聚类算法记为基于组合距离(hybrid distance)的 FCM 聚类算法(HDFCM)。

本节将上述 HDFCM 算法与已有的经典距离学习算法进行对比与分析。

3.1　实验设置

本文使用了 8 个来自 UCI 机器学习数据库的真实数据集。其中 4 个为二类数据集,其余 4 个为多类数据集。各个数据集的信息如表 1 所示。

表 1　实验中使用的数据集信息

Table 1　List of data sets

数据集	样本数	特征数	类别数
breast	683	10	2
sonar	208	60	2
wdbc	569	30	2
heart	270	12	2
wine	178	13	3
cmc	1 473	9	3
thyroid	215	5	3
segment	2 310	19	7

在数据集的选择上基于以下考虑:首先,这些数据集的特征数和类别数都各不相同。另外,这些数据集是机器学习研究中被广泛使用的基准数据集,因而具有代表性。最后,由于数据集均为真实数据集,因此可以检验算法在真实应用中是否可行。

文中所有实验均在 MATLAB 平台下进行,所有训练数据集和测试数据集均先归一化至 [0,1] 内。带有边信息的训练集将通过如下方法产生:首先,随机选取数据集的 10%组成一个子集。然后,根据子集中样本点带有的类标是否相同来生成约束对 $(\boldsymbol{x}_a^k, \boldsymbol{x}_b^k, y_k)$ 集合。其中,类标相同的成对约束为正约束对,反之为负约束对。将取个数相同的正负约束对组成训练集。

在组合距离分量的选择上,本文依据第 2 节的理论,在实验中选取如下 10 个距离度量进行组合:

$$d_1(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}} I (\boldsymbol{x}-\boldsymbol{y})$$

$$d_3(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}} \frac{3I}{\sigma^2}(\boldsymbol{x}-\boldsymbol{y})$$

$$d_4(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{d} |x_i - y_i|$$

$$d_5(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{d} \frac{|x_i - y_i|}{\sigma^2}$$

$$d_6(\boldsymbol{x},\boldsymbol{y}) = 1 - \mathrm{e}^{\frac{-3\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\sigma^2}}$$

$$d_7(\boldsymbol{x},\boldsymbol{y}) = 1 - \mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\sigma^2}}$$

$$d_8(\boldsymbol{x},\boldsymbol{y}) = 1 - \mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}{3\sigma^2}}$$

$$d_9(\boldsymbol{x},\boldsymbol{y}) = 1 - \mathrm{e}^{\frac{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}{5\sigma^2}}$$

$$d_{10}(\boldsymbol{x},\boldsymbol{y}) = 1 - \mathrm{e}^{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}$$

在使用组合距离进行聚类的算法中,本文将依据数据集的类别数给定聚类数目,初始隶属度矩阵随机生成。为了保证可比性,实验中所有的对比算法将使用相同的初始隶属度矩阵,训练集和其他参

数($m = 2, \varepsilon = 10^{-5}, T = 100, C = 10^{-5}$)。实验将重复每个聚类过程 20 次,实验结果取其均值。

为了评估聚类效果,采用一种类似 F_1-measure 的成对约束评价方法,评价参数包括:pairwise Precision,pairwise Recall 和 pairwise F_1, 定义为[2]

$$\text{Precision} = \frac{\#\text{TruePositive}}{\#\text{TruePositive} + \#\text{FalsePositive}}$$

$$\text{Recall} = \frac{\#\text{TruePositive}}{\#\text{TruePositive} + \#\text{FalseNegative}}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

式中: #TruePositive 为将正约束对预测为正约束对的个数, #FalsePositive 为将负约束对预测为正约束对的个数, #FalseNegative 为将正约束对预测为负约束对的个数。由于该评价方法的对象为约束对,因此不仅可以应用于二分类的评价,也可应用于多类分类的评价。

3.2 对比算法

本文使用了若干经典距离学习算法进行对比,包括:使用欧式距离的传统 FCM 算法(FCM),使用欧氏距离但含有约束条件的 K-均值聚类算法(C-Euc)[12],基于凸优化的全局距离学习算法(PGDM)[3]。

与本文提出的算法类似,C-Euc 算法也是一种利用边信息进行距离学习的半监督聚类算法,它在传统 K-均值算法的基础上加上成对约束,在这些约束的监督下进行聚类。C-Euc 算法在聚类的过程中要求每一次划分都满足已知的约束条件,每个样本在没有违反约束条件的情况下,被划分给最近的类,最终得到的聚类结果将满足所有的约束对信息[13]。

PGDM 算法由 Xing 等提出,是一种基于凸优化的全局距离度量学习算法。它将正约束对构成的集合记为 S ,负约束对构成的集合记为 D 。通过以下凸优化问题对距离矩阵 $\boldsymbol{A}$ 进行求解:

$$\min_{A} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in S} \| \boldsymbol{x}_i - \boldsymbol{x}_j \|_A^2$$

$$\text{s.t.} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in D} \| \boldsymbol{x}_i - \boldsymbol{x}_j \|_A \geqslant 1, \boldsymbol{A} \geqslant 0$$

式中: $\| \boldsymbol{x}_i, \boldsymbol{x}_j \|_A = \sqrt{(\boldsymbol{x}_i, \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{x}_i, \boldsymbol{x}_j)}$ 表示 2 个样本点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的距离。根据预期得到的矩阵 $\boldsymbol{A}$ 的不同将有不同的解法。如果期望得到对角形式的距离矩阵,可以通过牛顿法进行求解,本文将此算法记为 PGDM-Ad。如果期待得到全矩阵形式的距离矩阵,则可以通过梯度下降和逐次映射的方法进行求解,本文将此算法记为 PGDM-Af。为了保证对比性,在实验中本文将学习得到的距离矩阵和本文算法一样应用于 FCM 聚类算法中以评价。

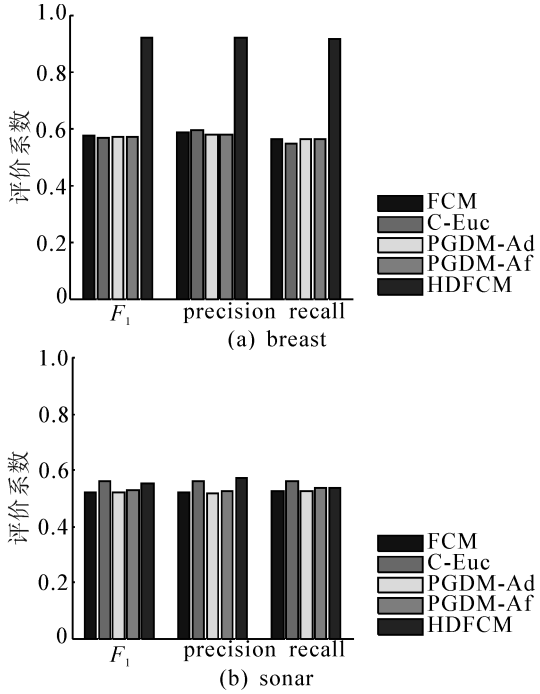
3.3 实验结果与分析

对于各个数据集,本文所提出算法与其他算法在 8 个数据集上的实验结果对比如图 1 所示,其中每一个子图的纵坐标表示了各个算法在相同参数下在该数据集上的聚类效果的评价指标均值,横坐标上的柱形分为 3 组,每一组分别表示 F_1、precsion 和 recall。每个颜色代表一个算法,从左至右分别为 FCM 算法[14], C-Euc 算法[12], PGDM-Ad 算法[3], PGDM-Af 算法[3]和 HDFCM 算法,数据集名称标注在图标题上。表 2 展示了本文算法相对于传统 FCM 算法聚类效果的提升率,提升率使用如下公式计算得到:

$$\text{提升率} = \frac{\text{HDFCM\_}F_1 - \text{FCM\_}F_1}{\text{FCM\_}F_1} \times 100\%$$

从图 1 可以看出,本文提出的算法在大部分数据集上获得了最好的表现。相对于其他距离学习算法而言,本文算法在 sonar 数据集和 cmc 数据集中虽未获得最好的表现,但是结合表 2 可以发现本文算法的聚类效果相对于传统 FCM 算法仍有一定的提升。由于本文使用的距离分量有限,因此对于不同的数据集不一定能拟合出最适合于该数据集的距离度量。此外,从表 2 可以观察到,本文算法在 breast 数据集和 wine 数据集上有相当卓越的表现。

结合图 1 和表 1 可以得出,本文算法不仅适用于 2 类数据集,对于多类数据集也有较好的聚类效果。比如,2 类数据集 breast,3 类数据集 wine,7 类数据集 segment 在聚类效果上均取得了 30%以上的提升。



(a) breast

(b) sonar

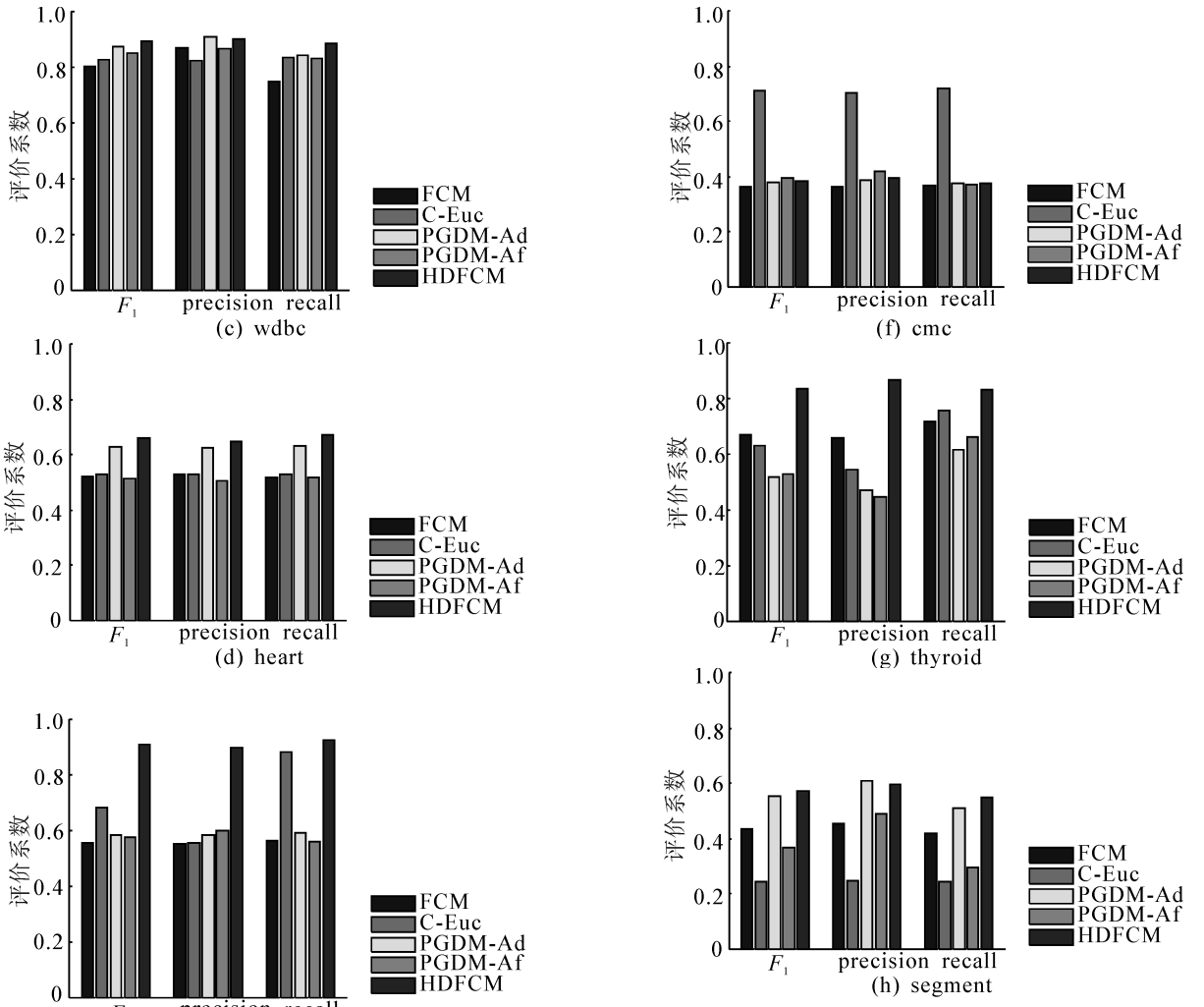


图 1　算法对比图

Fig.1　Clustering performance comparison

表 2　本文算法相对于传统 FCM 的提升率

Table 2　Upgrade rate of our algorithm

数据集	breast	sonar	wdbc	heart	wine	cmc	thyroid	segment
提升率/%	60.16	5.62	10.98	26.28	64.24	5.33	24.71	31.29

由于传统 FCM 算法使用的是欧式距离,且其为无监督聚类算法,因此在应用的过程中不一定适合所有类型的数据集。而 C-Euc 算法虽然引入了数据集的边信息,但是其使用的距离度量仍然为欧氏距离,因此在使用的时候也具有局限性。PGDM 在引入了边信息的基础上学习出了新的距离度量,但是该距离函数仍是在马氏距离定义框架下的距离度量,属于线性的距离学习方法。本文提出的算法不仅引入了数据集的边信息,而且组合了预设的多种形式的距离度量,学习得到一个非线性的距离度量,使其对于数据集有较好的适应性。上述实验可以证明本文算法的有效性。

4　结束语

本文提出了一种基于线性组合的混合距离学习算法。该算法构建了一个由若干候选距离线性组合而成的距离目标函数,利用数据集的边信息学习得到各候选距离对应权值,从而得到新的距离函数。本文将学习得到的距离函数应用于模糊 C 均值算法中以构成一个半监督聚类算法。通过使用 UCI 真实数据集将该半监督聚类算法的聚类效果与其他距离学习算法进行对比,证明了本文算法的有效性。

参考文献:

[1]王骏, 王士同. 基于混合距离学习的双指数模糊 C 均值算法[J]. 软件学报, 2010, 21(8): 1878-1888.

WANG Jun, WANG Shitong. Double indices FCM algorithm based on hybrid distance metric learning[J]. Journal of Software, 2010, 21(8): 1878-1888.

[2]WU Lei, HOI S C H, JIN Rong, et al. Learning Bregman distance functions for semi-supervised clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 478-491.

[3]XING E P, NG A Y, JORDAN M I, et al. Distance metric learning with application to clustering with side information[C]//Advances in Neural Information Processing Systems. Vancouver, Canada, 2002: 521-528.

[4]HOI S C H, LIU Wei, LYU M R, et al. Learning distance metrics with contextual constraints for image retrieval[C]// Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, America, 2006, 2: 2072-2078.

[5]GOLDBERGER J, HINTON G, ROWEIS S, et al. Neighborhood component analysis[C]//Advances in Neural Information Processing Systems. Cambridge, United Kingdom, 2005: 451-458.

[6]WEINBERGER K Q, BLITZER J C, SAUL L K. Distance metric learning for large margin nearest neighbor classification[C]//Advances in Neural Information Processing Systems. Cambridge, United Kingdom, 2006: 1473-1480.

[7]王骏, 王士同, 邓赵红. 特征加权距离与软子空间学习相结合的文本聚类新方法[J]. 计算机学报, 2012, 35(8): 1655-1665.
WANG Jun, WANG Shitong, DENG Zhaohong. A novel text clustering algorithm based on feature weighting distance and soft subspace learning[J]. Chinese Journal of Computers, 2012, 35(8): 1655-1665.

[8]WU K L, YANG M S. Alternative c-means clustering algorithms[J]. Pattern Recognition, 2002, 35(10): 2267-2278.

[9]CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[10]TSANG I W H, KWOK J T Y, ZURADA J A. Generalized Core Vector Machines[J]. IEEE Transaction on Neural Networks, 2006, 17(5): 1126-1140.

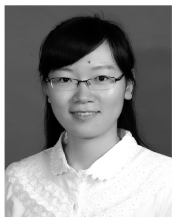
[11]MEI Jianping, CHEN Lihui. Fuzzy clustering with weighted medoids for relational data[J]. Pattern Recognition, 2010, 43(5): 1964-1974.

[12]WAGSTAFF K, CARDIE C, ROGERS S, et al. Constrained k-means clustering with background knowledge[C]// BAR-HILLEL A, HERTZ T, SHENTAL N, et al. Proceedings of the Eighteenth International Conference on Machine Learning. Williamstown, Australia, 2001: 577-584.

[13]COVÕES T F, HRUSCHKA E R, GHOSH J. A study of k-means-based algorithms for constrained clustering[J]. Intelligent Data Analysis, 2013, 17(3): 485-505.

[14]BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum Press, 1981: 56-57.

作者简介：



郭瑛洁，女，1991 生，硕士研究生，主要研究方向为人工智能、模式识别。

王士同，男，1964 生，教授，博士生导师，主要研究方向为人工智能、模式识别和生物信息。

许小龙，男，1989 生，硕士研究生，主要研究方向为人工智能、模式识别。