文章编号:1001 - 9081(2007)03 - 0553 - 03

基于模糊核学习矢量量化的 Sammon 非线性映射算法

晋良念,欧阳缮, 李民政
(桂林电子科技大学 信息与通信学院,广西 桂林 541004)
(jing@guet.edu.cn; jinglingling5653@sina.com.cn)

摘 要:提出了一种基于可靠稳定的模糊核学习矢量量化(FKLVQ)聚类的 Sammon 非线性映射新算法。该方法通过 Mercer 核,将数据空间映射到高维特征空间,并在此特征空间上进行 FKLVQ 学习获取数据空间有效且稳定的聚类权矢量,然后在特征空间和输出空间上仅针对各空间的数据样本和它们各自的聚类权矢量进行 Sammon 非线性核映射。这样既降低了计算的复杂度,又使数据空间和输出空间上数据点与聚类中心间的距离信息保持相似。仿真结果验证了该方法的可靠性和稳定性。

关键词:非线性映射;Sammon 投影;距离保持性;计算复杂度;模糊核;学习矢量量化
中图分类号: TP311.13 **文献标识码**:A

Algorithm for Sammon's nonlinear mapping based on fuzzy kernel learning vector quantization

JIN Liang-nian, OU Yang-shan, LI Min-zheng
(*Information and Communication College, Guilin University of Electronic Technology, Guilin Guangxi* 541004, *China*)

Abstract: An new algorithm for Sammon's nonlinear kernel mapping based on reliable and stable fuzzy kernel learning vector quantization was presented. The data space was mapped to high dimension feature space with Mercer kernel function, and fuzzy kernel learning vector quantization (FKLVQ) was done on the feature space to obtain the effective and stable clustering weight vectors. Finally Sammon's nonlinear kernel mapping only for the data points and the clusters was executed on the output space and the feature space, thus reducing computational complexity and preserving the distance resemblance between the clusters and the data points from the data space to the output space. Simulation results demonstrate the reliability and stability of the proposed algorithm.

Key words: nonlinear mapping; Sammon's projection; distance preservation; computational complexity; fuzzy kernel; Learning Vector Quantization (LVQ)

0 引言

数据降维映射是数据投影、数据挖掘、可视化或聚类分析的基础。然而,当今数据量及其维数的急剧膨胀使得数据降维面临着巨大挑战。由于人眼视觉对二维或三维空间上数据的分布特性有绝佳的分辨能力,所以通过投影将高维空间的数据映射到尽可能保持原空间数据的某种内在关系的低维空间上进行适当的区分和分类是聚类分析的重要途径之一。Sammon 映射就是"几何图像降维"投影法,它通过非线性变换,在低维空间上直观、形象地展现原数据间的结构信息,使得人们能够在低维空间上看到一些高维样本点相互关系的近似图像。但是它存在计算复杂度大,对初值的设定较敏感以及易陷入局部极值等缺点[1]。映射初值的优化设定[2]以及满足 AGW 条件的线性搜索[3]能较好地解决初值敏感性和局部极值问题。本文重点针对计算复杂度较大的缺点进行讨论。混合 FCM 聚类的 Sammon 算法[4]能够较好地解决计算复杂度的问题,它的思想是首先通过模糊 C 均值(Fuzzy C-Mean, FCM)算法得到各空间的聚类中心,然后在数据空间和输出空间上仅针对各自空间的数据点和它们各自的聚类中心进行 Sammon 非线性映射,以尽可能地保证两空间数据点的距离相似性,算法对线性可分数据集的映射接近 Sammon 映射,从而说明算法可行。但是,FCM 算法聚类的缺点极易使 Sammon 映射扭曲低维输出空间的距离信息以及混合 FCM-Sammon 算法对许多现实数据集表现出较差的可靠性和推广能力。为此,本文提出一种 Sammon 非线性映射的新方法——基于模糊核学习矢量量化(Fuzzy Kernel Learning Vector Quantization, FKLVQ)算法,它应用核思想为原数据空间诱导出一类异于欧氏距离度量的新的灵活的距离度量以提高可靠性和推广能力,应用 FKLVQ 算法以提高聚类中心的有效性和稳定性。

1 模糊核学习矢量量化(FKLVQ)算法

由于 FCM 具有聚类中心的随机初始化导致最终迭代结果的不稳定,权重指数的选择直接影响聚类有效性,数据点模糊隶属各类中心的非独立性导致算法对噪声数据非常敏感以及每一步迭代均需对整个数据集进行计算导致较大计算量等等缺点[5],所以混合 FCM 的 Sammon 算法[4]处理大容量的多维数据难以取得理想的效果。近年来,Karayiannis 和 Pai 提

收稿日期:2006 - 09 - 15 **基金项目**:国家自然科学基金资助项目(60172011)
作者简介:晋良念(1974 -),男,四川简阳人,讲师,博士研究生,主要研究方向:数据挖掘、自适应信号处理、神经网络; 欧阳缮(1960 -),男,江西安福人,教授, 博士生导师, 主要研究方向:自适应信号处理、通信信号处理、神经网络; 李民政(1972 -),男,甘肃兰州人,讲师,博士研究生,主要研究方向:信息科学、多媒体信息处理.

出的模糊学习矢量量化(FLVQ)[6,7]是基于“winner-take-most”竞争策略的随机梯度学习算法,它在一定程度上克服了FCM算法的缺点,但是FLVQ的稳定性受到学习率的影响,必须通过调整学习率来保证算法稳定地收敛,否则学习率过大导致算法不收敛而远离训练集,学习率过小导致算法在初始值附近徘徊。我们发现,FLVQ出现该问题的原因是:每个非获胜端的隶属函数μ_{ik}均包含获胜端的贡献,使得获胜端对目标函数的贡献扩大到聚类数目c的倍数,所以获胜端码向量的调整随c增加引起发散,这样只能改变初始学习速率以满足算法有效的必要条件,从而保证算法收敛。尽管文献[8]提出了LVQ的一般模型,其定义的隶属函数克服了对c获胜端码向量调整的影响,但是非获胜型端码向量随c的增加其调整量极小,几乎在初值徘徊,所以该学习规则并非符合FLVQ算法的“winner-take-most”竞争策略。另外,这些算法对c较小且线性可分的数据集的聚类能取得满意的效果,但是对c较大或一些线性不可分但非线性可分的数据集却表现出较差的结果。为此,将原FLVQ算法进行改进,把获胜端i^*的隶属度函数扩展c倍并结合原空间核的思想,得到模糊核学习矢量量化(FKLVQ)的目标函数的定义:

$$E(k) = \sum_{i=1}^{c} \mu_{ik} \| \phi(\boldsymbol{X}_k) - \phi(\boldsymbol{W}_{ik}) \|^2 \tag{1}$$

$$\mu_{ik} = \begin{cases} c \cdot u_{i*k}, & i = i^* \\ u\left(\dfrac{\| \phi(\boldsymbol{X}_k) - \phi(\boldsymbol{W}_{i*k}) \|^2}{\| \phi(\boldsymbol{X}_k) - \phi(\boldsymbol{W}_{ik}) \|^2}\right) = u(z_{ik}), & i \neq i^* \end{cases} \tag{2}$$

其中,获胜端隶属度u_{i*k}是关于变量集$\{z_{1k}, z_{2k}, \cdots, z_{ck}\}$的函数;向量$\boldsymbol{X}_k$是第$k$节拍从数据样本集$\boldsymbol{X}_p$,$p \in \{1, 2, \cdots, M\}$中依次取出的数据样本;向量$\boldsymbol{W}_{jk}$,$j \in \{1, 2, \cdots, c\}$是第$k$节拍的聚类中心点;$i^*$为输入$\boldsymbol{X}_k$时的获胜端编号;$i$为输入$\boldsymbol{X}_k$时的非获胜端编号;$\phi$是核映射,其映射得到的特征空间中的点积用数据空间的核来表示,即$K(\boldsymbol{X}_i, \boldsymbol{X}_k) = <\phi(\boldsymbol{X}_i), \phi(\boldsymbol{X}_k)>$。因为高斯核函数其核函数对应的特征空间$H$是无穷维的,从而有限的数据样本在该特征空间肯定是线性可分的,所以选用$K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\| \boldsymbol{x} - \boldsymbol{y} \|^2 / 2\sigma_1^2)$,那么$z_{ik} = [2 - 2 * K(\boldsymbol{X}_k, \boldsymbol{W}_{i*k})] / [2 - 2 * K(\boldsymbol{X}_k, \boldsymbol{W}_{ik})]$,$i \neq i^*$。

将$E(k)$关于非获胜端$\boldsymbol{W}_{ik}$和获胜端$\boldsymbol{W}_{i*k}$求导,并在$\dfrac{\partial u(z_{ik})}{\partial z_{ik}} + z_{ik} \dfrac{\partial (cu_{i*k})}{\partial z_{ik}} = 0$的约束条件下(目的是保证算法更新的尺度函数为模糊隶属函数),导出FKLVQ算法的学习规则为:

$$\boldsymbol{W}_{i*k+1} = \boldsymbol{W}_{i*k} + \varepsilon_k \cdot K(\boldsymbol{X}_k, \boldsymbol{W}_{i*k}) \cdot (\boldsymbol{X}_k - \boldsymbol{W}_{i*k}) \cdot u_{i*k} \tag{3}$$

$$\boldsymbol{W}_{ik+1} = \boldsymbol{W}_{ik} + \varepsilon_k \cdot K(\boldsymbol{X}_k, \boldsymbol{W}_{ik}) \cdot (\boldsymbol{X}_k - \boldsymbol{W}_{ik}) \cdot u(z_{ik}) \tag{4}$$

其中,$i = 1, 2, \cdots, c$,$i \neq i^*$。

由于获胜端隶属度$u_{i*k} \in [0, 1]$是变量集$\{z_{1k}, z_{2k}, \cdots, z_{ck}\}$的函数,所以令$u_{i*k} = \dfrac{1}{c} + \sum_{i \neq i^*} \dfrac{1}{c} g(z_{ik})$(这里,$g(z_{ik})$是只与$z_{ik}$有关的干扰函数,在[0,1]上递减,并且满足$0 < g(z_{ik}) \leqslant 1$)。根据$\dfrac{\partial u(z_{ik})}{\partial z_{ik}} + z_{ik} \dfrac{\partial (cu_{i*k})}{\partial z_{ik}} = 0$约束条件得到干扰函数$g(z_{ik})$和非获胜端隶属度$u(z_{ik})$之间的关系,即$u(z_{ik}) = -z_{ik} g(z_{ik}) + \int_0^{z_{ik}} g(x)\, \mathrm{d}x$。表1给出了FKLVQ算法三种类型的干扰函数$g(z)$和隶属度函数$u(z)$。另外,为了满足“winner-all-most”的竞争策略,使得算法的聚类性能稳定,要求$0 \leqslant u(z_{ik}) \leqslant u_{i*k} \leqslant 1$,这致使参数$\alpha, \beta, \gamma$的选择必满足一定取值范围。

表1 FKLVQ算法的干扰函数$g(z)$和隶属度函数u_i

类型	$g(z)$	$u(z)$
类型1	$(1 + \alpha z)^{-2}$	$\alpha z^2 (1 + \alpha z)^{-2}$
类型2	$(1 - \beta z) \exp(-\beta z)$	$\beta z^2 \exp(-\beta z)$
类型3	$1 - 2\gamma z$	γz^2

2 基于FKLVQ的Sammon核算法

与FCM相比,因FKLVQ算法具有较高的可靠性、稳定性以及较好的推广能力使得聚类效果稳定有效。当FKLVQ获取有效的聚类中心后,在特征空间和输出空间上仅针对各空间的数据样本和它们各自的聚类权矢量进行Sammon非线性核映射。令$\boldsymbol{W}_i$,$i \in \{1, 2, \cdots, c\}$为在特征空间上进行FKLVQ算法学习得到的数据空间的聚类权向量,$\boldsymbol{X}_k \in R^L$,$k \in \{1, 2, \cdots, M\}$为数据空间的数据点,$\boldsymbol{Z}_i$,$i \in \{1, 2, \cdots, c\}$为输出空间的聚类权向量,$\boldsymbol{y}_k \in R^D$,$k \in \{1, 2, \cdots, M\}$为输出空间的数据点,$d_{ik}^*$为原数据空间的数据点与其聚类中心在均映射到特征空间后的欧氏距离,即$d_{ik}^* = \sqrt{K(\boldsymbol{W}_i, \boldsymbol{W}_i) + K(\boldsymbol{X}_k, \boldsymbol{X}_k) - 2K(\boldsymbol{W}_i, \boldsymbol{X}_k)}$,$d_{ik}$为输出空间上的数据点与其聚类权向量的欧氏距离,即$d_{ik} = \| \boldsymbol{Z}_i - \boldsymbol{y}_k \|$,$\lambda$为分辨率参数,目的是调整输出空间上数据间的分辨距离,若λ较大,分辨率较高,反之分辨率较低。基于FKLVQ的Sammon非线性核算法的目标函数可定义为:

$$E = \frac{1}{\sum_{i=1}^{c} \sum_{k=1}^{M} \lambda d_{ik}^*} \sum_{i=1}^{c} \sum_{k=1}^{M} \frac{(\lambda d_{ik}^* - d_{ik})^2}{\lambda d_{ik}^*} \tag{5}$$

为保证输出空间聚类权$\boldsymbol{Z}_i$和特征空间聚类权$\boldsymbol{W}_i$的一致性,需应用FKLVQ学习算法获得聚类权$\boldsymbol{W}_i$稳定解所对应的隶属度函数和高斯核函数来更新$\boldsymbol{Z}_i$,所以:

$$\boldsymbol{Z}_i = \sum_{k=1}^{M} K(\boldsymbol{X}_k, \boldsymbol{W}_{ik}) u_{ik} \boldsymbol{y}_k \Big/ \sum_{k=1}^{M} K(\boldsymbol{X}_k, \boldsymbol{W}_{ik}) u_{ik}, \quad i = 1, 2, \cdots, c \tag{6}$$

应用牛顿迭代算法,令$s = \sum_{i=1}^{c} \sum_{k=1}^{M} \lambda d_{ik}^*$,将$E$关于$y_{kq}$求导,得到:

$$\frac{\partial E}{\partial y_{kq}} = -\frac{2}{s} \sum_{j=1}^{c} \left(\frac{\lambda d_{kj}^* - d_{kj}}{\lambda d_{kj}^* d_{kj}}\right)(y_{kq} - z_{jq}), \quad k = 1, 2, \cdots, M,\ q = 1, 2, \cdots, D \tag{7}$$

$$\frac{\partial^2 E}{\partial y_{kq}^2} = -\frac{2}{s} \sum_{j=1}^{c} \left[\frac{1}{d_{kj}} - \frac{1}{\lambda d_{kj}^*} - \frac{(y_{kq} - y_{jq})^2}{d_{kj}^3}\right]$$

$$y_{kq}(m+1) = y_{kq}(m) - \alpha \cdot \left|\frac{\partial^2 E}{\partial y_{kq}^2}\right|^{-1} \frac{\partial E}{\partial y_{kq}}$$

$$k = 1, 2, \cdots, M,\ q = 1, 2, \cdots, D \tag{8}$$

α的取值与原始Sammon算法的迭代步长一致,通常取$\alpha = 0.2 \sim 0.4$。当然,为避免陷入局部极值,也可参考满足AGW条件的线性搜索方法[3]选择最佳α。另外根据上述的分析,基于FKLVQ的Sammon算法每次迭代的计算复杂度和FCM-Sammon算法的计算复杂度均为$O(M \cdot c)$,远小于Sammon映射的$O(M^2)$。下面给出基于FKLVQ的Sammon非线性映射算法的具体步骤:

1)使用FKLVQ算法对原数据进行聚类,得到隶属度矩阵$\boldsymbol{U} = \{u_{ik}\}$和聚类权矢量$\boldsymbol{W}_i$,$i = 1, 2, \cdots, c$;

2) 随机初始化 $\boldsymbol{y}_j \in R^p$, $j = 1, 2, \cdots, M$,用(6)式计算初始化 $\boldsymbol{y}_j$ 的聚类权 $\boldsymbol{Z}_i$, $i = 1, 2, \cdots, c$;

3) 用(7)和(8)式计算 $\boldsymbol{y}_j$, $j = 1, 2, \cdots, M$;

4) 用(6)式计算聚类权矢量 $\boldsymbol{Z}_i$, $i = 1, 2, \cdots, c$;

5) 循环执行 3) ~ 4),直至迭代最大次数或满足代价目标函数误差限的条件。

3　仿真结果

表 2　三种算法的 *RSD* 比较

数据样本集	基于 FKLVQ 的 Sammon 算法	Sammon 算法	基于 FCM 的 Sammon 算法
2D 合成数据集	0.0292	0.0436	0.0098
Iris 数据集	0.2440	0.2460	0.3024

表 3　三种算法的代价目标函数值比较

数据样本集	基于 FKLVQ 的 Sammon 算法	Sammon 算法	基于 FCM 的 Sammon 算法
2D 合成数据集	0.000196	0.000116	0.000007
Iris 数据集	0.006084	0.006320	0.004778

为了比较 Sammon 算法、基于 FCM 聚类的 Sammon 算法以及基于 FKLVQ 聚类的 Sammon 算法间性能,分别用 2D 非线性合成数据样本集和著名的 Iris 样本集进行测试。其中 2D 非线性可分数据样本集由满足均匀分布且半径均值分别取 1,6,10 的随机环形数据组成,各类数据量分别取 100。为了定量比较这些算法间的性能,引入表征将输入空间上任一数据点到其邻域内各数据点的距离分别与输出空间上对应的距离相比所得比值分布的偏离方差(*RSD*) 以刻画两空间上数据间距离的相似程度。表 2 给出了这些算法的 *RSD* 值,表 3 给出了各算法的代价目标函数值 E。根据表 2 的结果,基于 FKLVQ 的 Sammon 算法的 *RSD* 值与 Sammon 映射的 *RSD* 值接近且表现稳定,但是基于 FCM 的 Sammon 算法的 *RSD* 值均偏离 Sammon 映射的 *RSD* 值,表现不稳定,此结果证实了基于 FKLVQ 的 Sammon 算法的稳定性优于基于 FCM 的 Sammon 算法。另外,基于 FKLVQ 的 Sammon 算法的 值均稍小于 Sammon 映射,说明引入核的混合 FKLVQ-Sammon 算法具有较好的紧致性和可靠性。根据表 3,基于 FKLVQ 的 Sammon 算法的 E 值均与 Sammon 映射的 E 值接近,但是基于 FCM 的 Sammon 算法的 E 值对初始值的设定较敏感, E 值均偏离 Sammon 映射的 E 值,表现不够稳定,此结果也证实了基于 FKLVQ 的 Sammon 算法在稳定性能上优于基于 FCM 的 Sammon 算法。综上所述,基于 FKLVQ 聚类的 Sammon 算法是可靠、有效而且比较稳定的。

4　结语

针对大容量数据情况下 Sammon 映射计算复杂度较大以及混合 FCM-Sammon 算法的不稳定性,提出了一种基于有效且稳定的 FKLVQ 聚类的 Sammon 映射新算法。它继承了混合 FCM-Sammon 算法能降低计算复杂度的优点,克服了混合 FCM-Sammon 算法结果不稳定的缺点,并应用核思想从原数据空间诱导出一类异于欧氏距离的新的距离度量,提高了算法的可靠性和推广能力。仿真结果在适宜的参数(如 σ 和 β 等)下验证了提出算法的有效性和可行性,其性能优于 Sammon 算法和混合 FCM 的 Sammon 算法。因 FKLVQ 聚类结果对映射极为重要,其参数的选择直接影响到最终的映射性能,所以如何根据 FKLVQ 的收敛性和稳定性选择较佳参数的问题还有待深入分析。

参考文献:

[1] SAMMON JW . A Nonlinear Mapping for Data Structure Analysis [J]. IEEE Transactions on Computers, 1969,18(5), 401 -409.

[2] MOTVILAS AM. Optimal Initial Conditions for Nonlinear Mapping of Multidimensional Signals[J]. ELEKTRONIKA IR ELEKTROTECHNIKA,2005,57(1):24 -27.

[3] DEODHARE D , KESHEOREY A , SHARMA A . An Improved Sammon's Nonlinear Mapping Algorithm[A]. Proceedings of the International Conference on Cognition and Recognition[C]. 2005.74 -82.

[4] KOVáCS A, ABONYI J. Visualization of Fuzzy Clustering Results by Modified Sammon Mapping[A]. ELEKTRONIKA IR ELEKTROTECHNIKA,2005,57(1):45 -48.

[5] 于剑. 论模糊 C 均值算法的模糊指标[J]. 计算机学报,2003,26(8):968 -973.

[6] KARAYIANNIS NB, PAI P-I. Fuzzy Algorithms for Learning Vector Quantization[J]. IEEE Transactions on Neural Networks, 1996,7(5):1196 -1211.

[7] KARAYIANNIS NB. A Methodology for Constructing Fuzzy Algorithms for Learning Vector Quantization[J]. IEEE Transactions on Neural Networks, 1997,8(3):505 -518.

[8] 张志华,郑南宁, 王天树. 学习矢量量化的软竞争算法[J]. 软件学报, 2002, 13(5):980 -987.

(上接第 552 页)

[5] 常犁云,王国胤,吴渝. 一种 Rough Set 理论的属性约简及规则提取方法[J]. 软件学报,1999,10(11):1206 -1211.

[6] 苗夺谦,胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999,36(6): 681 -684.

[7] 王珏,王任,苗夺谦,等. 基于 Rough Set 理论的"数据浓缩"[J]. 计算机学报,1998,21(5):393 -400.

[8] 叶东毅. Jelonek 属性约简算法的一个改进[J]. 电子学报,2000,28(12):81 - 82.

[9] JAKUB W. Finding mininmal reducts using genetic algorithm, ICS Research Report 16/95[R]. Warsaw University of Technology,1995.

[10] 李订芳,章文,李贵斌,等. 基于可行域的遗传约简算法[J]. 小型微型计算机系统,2006,27(2):312 -315.

[11] DAI J-H,LI Y-X. Heuristic genetic algorithm for minimal reduction decision system based on rough set theory[A]. Proceedings of 2002 International Conference on Machine Learning and Cybernetics[C]. 2002,2.4 -5.

[12] KENNEDY J,EBERHART RC. A discrete binary version of the particle swarm algorithm[A]. Proceedings of the 1997 Conference on Systems, Man, and Cybernetics[C]. Piscataway: IEEE Press, 1997.4104 -4109.

[13] 王磊,潘近,焦李成. 免疫算法[J]. 电子学报, 2000, 28(7): 74 -78.

[14] 焦李成,杜海峰. 人工免疫系统进展与展望[J]. 电子学报. 2003,31(10): 1540 -1548.

[15] 高鹰,谢胜利. 免疫粒子群优化算法[J]. 计算机工程与应用, 2004, 40(6): 4 -7.

[16] 叶东毅,廖建坤. 基于粒子群优化的最小属性约简算法[A]. 第11届中国人工智能大会论文集[C]. 北京:北京邮电大学出版社, 2005.728 -732.