优化 KPCA 特征提取下的 FCM 算法研究

蔡静颖,张 永,张凤梅,谢福鼎
CAI Jing-ying,ZHANG Yong,ZHANG Feng-mei,XIE Fu-ding

辽宁师范大学 计算机系,辽宁 大连 116081
Department of Computer,Liaoning Normal University,Dalian,Liaoning 116081,China
E-mail:ayong_zh@163.com

CAI Jing-ying,ZHANG Yong,ZHANG Feng-mei,et al.Fuzzy C-Mean algorithm based on optimized KPCA feature extraction.Computer Engineering and Applications,2009,45(32):38-40.

Abstract: Kernel PCA method extracts feature from large samples and high dimension data sets,combining CA to select optimized kernel function or near optimized kernel function.FCM based on the method not only effectively extracts the nonlinear information from the samples but also reduces dimension.Experiment shows its better clustering result and less train time.
Key words: kernel principle component analysis;cultural algorithm;fuzzy clustering

摘 要:利用核函数主元分析(KPCA)方法对大样本、高维数据进行特征提取预处理,并结合文化算法(CA)选择最优或接近最优的核函数,将其用于模糊 C 均值(FCM)聚类中,不但有效地提取了样本的非线性信息,而且使样本维数得到约简。实验表明该方法具有较好的聚类效果和更少的训练时间。
关键词:核函数主元分析;文化算法;模糊聚类
DOI:10.3778/j.issn.1002-8331.2009.32.012 **文章编号**:1002-8331(2009)32-0038-03 **文献标识码**:A **中图分类号**:TP391

1 引言

模糊 C-均值算法(Fuzzy C-Means,FCM)是由 Bezdek[1]于 1981 年提出的基于模糊集合理论的聚类算法,该算法是目前应用最为广泛的聚类算法之一。传统的模糊 C 均值算法是基于欧式距离的,在高维数据集中,传统的欧几里德密度定义(单位体积中点的个数)变得没有意义,这使得模糊 C 均值算法在处理高维数据集时效果很差,处理该问题的一种方法是使用维归约技术,其中主成分分析(Principle Component Analysis,PCA)就是一种最常用的维归约线性代数技术。由于线性主元分析方法对非线性特征不能有效提取,在 PCA 中引入核函数,称为核函数主元分析(Kernel Principle Component Analysis,KPCA)[2],可以有效地提取输入数据的非线性信息。

核主元分析(KPCA)是一种输入输出特征非线性变换技术,在特征空间进行主元分析,以获得一种对故障识别能力最优的特征变换,该方法已经被广泛地应用于模式识别、数据分类与聚类,化工建模等领域的数据预处理中。文献[3]将 KPCA 应用于人脸识别中,取得了较好效果;文献[4]表明,经过 KPCA 特征提取之后的 SVM 具有更好的分类精度和分类速度;文献[5]将 KPCA 与 SVM 结合应用于化工建模。但核主元分析法存在如何根据具体问题选择最优核函数及参数问题,该文根据文献[6]提出的基于文化算法的 KPCA 进行特征提取,从而进行核函数的最优或近似最优选择。

文化算法(Cultural Algorithms,CA)是一种新的进化算法[7],是将个人的以往经验保存于其中的知识库,新的个人可以在知识库中学到他没有直接经历的经验知识[8]。文化算法对许多典型问题具有良好的优化性能,目前文化算法已应用于数据挖掘、资源调度、函数优化、欺骗探测、遗传规则、动态环境建模等领域[9-10],而国内刚刚开始关注文化算法的研究。

该文将 KPCA 与 CA 两者有效地结合起来,提高核函数的优选,并将其应用于模糊 C 均值的高维数据预处理中。实验验证其算法较之传统的模糊 C 均值算法训练时间降低,聚类精度提高。

2 模糊 C-均值聚类方法

设 $X=\{x_1,x_2,\cdots,x_n\}$为 n 元数据集合,$x_i\in R^s$。FCM 聚类方法就是把 X 划分为 c 个子集 $S_1,S_2,\cdots,S_c$,若用 $a_1,a_2,\cdots,a_c$ 表示

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.10771092);辽宁省博士启动基金(the Liaoning Doctoral Research Foundation of China under Grant No.20081079);辽宁省教育厅科学技术研究项目(the Scientific Research Project of Liaoning Education Department of China under Grant No.2008347)。
作者简介:蔡静颖(1975-),女,讲师,主要研究方向为人工智能,数据挖掘;张永(1975-),通讯作者,男,博士,副教授,主要研究领域为机器学习,智能计算,可信计算;张凤梅(1970-),女,讲师,主要研究领域为人工智能;谢福鼎(1965-),男,博士,教授,主要研究方向为人工智能,数据挖掘。
收稿日期:2008-11-27 **修回日期**:2009-02-03

这 c 个子集的聚类中心，u_{ij} 表示元素 x_j 对 S_i 的隶属度，则 FCM 算法的优化目标函数为：

$$J_{\mathrm{FCM}}^{m}(\boldsymbol{U},\boldsymbol{A},X)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2}=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\parallel x_j-a_i\parallel \tag{1}$$

约束条件：

$$\sum_{i=1}^{c}u_{ij}=1,1\leqslant j\leqslant n,u_{ij}\geqslant 0,1\leqslant i\leqslant c,1\leqslant j\leqslant n \tag{2}$$

这里 $\boldsymbol{U}=\{u_{ij}\}$ 为 $c\times n$ 矩阵，$\boldsymbol{A}=\{a_1,a_2,\cdots,a_c\}$ 为 $s\times c$ 矩阵，d_{ij} 为 x_j 与 a_i 的距离，经典的 FCM 算法里使用欧氏距离。m 为大于 1 的模糊指数，控制分类矩阵 $\boldsymbol{U}$ 的模糊程度，m 越大，分类的模糊程度越高，在实际应用中 m 最佳范围为(1.5，2.5)，推荐使用 m=2。FCM 算法是使目标函数最小化的迭代收敛过程。在迭代求解 J_{FCM} 的最小值时，u_{ij} 是按 Lagrange 乘数法得到的：

$$a_i=\frac{\sum_{j=1}^{n}(u_{ij})^{m}x_j}{\sum_{j=1}^{n}(u_{ij})^{m}} \tag{3}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}} \tag{4}$$

可以看出，FCM 算法就是反复修改聚类中心矩阵和隶属度矩阵的分类过程。

3　KPCA 原理及其优化

3.1　KPCA 原理

KPCA 方法并不是直接计算样本协方差矩阵的特征向量，而是将其转化为求核矩阵的特征值和特征向量问题，从而避免了在整个特征空间上求特征向量。与其他非线性特征提取方法相比，它不需要解非线性优化问题。

KPCA 先对任一样本 x 进行非线性变换 $\boldsymbol{\Phi}(x)$，$\boldsymbol{\Phi}(x)$ 将 x 映射到高维空间中。对于新的样本空间，此时的协方差矩阵为：

$$\boldsymbol{C}=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{\Phi}(x_i)\boldsymbol{\Phi}(x_i)^{\mathrm{T}} \tag{5}$$

主成分分析的任务即求出特征值 λ 和特征向量 $\boldsymbol{v}$。由 $\lambda\boldsymbol{v}=\boldsymbol{C}\boldsymbol{v}$ 可得，当 $\lambda\neq 0$ 时，$\boldsymbol{v}$ 在 $\boldsymbol{\Phi}(x_i)(i=1,2,\cdots,M)$ 张成的空间中，即存在 $\alpha_i(i=1,2,\cdots,M)$，满足

$$\boldsymbol{v}=\sum_{i=1}^{M}\alpha_i\boldsymbol{\Phi}(x_i) \tag{6}$$

在 $\lambda\boldsymbol{v}=\boldsymbol{C}\boldsymbol{v}$ 两边同时与 $\boldsymbol{\Phi}(x_k)$ 做内积得

$$\lambda\langle\boldsymbol{\Phi}(x_k),\boldsymbol{v}\rangle=\langle\boldsymbol{\Phi}(x_k),\boldsymbol{C}\boldsymbol{v}\rangle,k=1,2,\cdots,M \tag{7}$$

将式(5)和式(6)代入式(7)后转化为如下问题：

$$M\lambda\boldsymbol{\alpha}=\boldsymbol{K}\boldsymbol{\alpha},\boldsymbol{K}=(\kappa(x_i,x_j))_{ij} \tag{8}$$

其中，$\boldsymbol{K}$ 是核矩阵，$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_M)'$。可得 $M\lambda$ 和 $\boldsymbol{\alpha}$ 是对应于 $\boldsymbol{K}$ 的特征值和特征向量。假设对应于大于 0 的特征值的特征向量分别为 $\boldsymbol{\alpha}^{p},\boldsymbol{\alpha}^{p+1},\cdots,\boldsymbol{\alpha}^{M}$，为了满足 $\langle\boldsymbol{v}^{r},\boldsymbol{v}^{r}\rangle=1$，取 $\boldsymbol{\alpha}^{r}$ 使得 $M\lambda\langle\boldsymbol{\alpha}^{r},\boldsymbol{\alpha}^{r}\rangle=1$，则样本 $\boldsymbol{\Phi}(x)$ 在 $\boldsymbol{v}^{r}$ 上的投影为：

$$g_r(x)=\langle\boldsymbol{v}^{r},\boldsymbol{\Phi}(x)\rangle=\sum_{i=1}^{M}\alpha_i^{r}k(x_i,x) \tag{9}$$

其中，$r=p,p+1,\cdots,M$。$g_r(x)$ 为对应于 $\boldsymbol{\Phi}(x)$ 的非线性主元分量，将所有投影值形成的向量 $(g_1(x),g_2(x),\cdots,g_l(x))'$ 作为样本 x 的新特征。

由此可见，KPCA 的分析计算过程可分为 3 步：

(1)计算出矩阵 $\boldsymbol{K}$；

(2)计算它的特征向量并在 $\boldsymbol{\Phi}(x)$ 空间进行标准化；

(3)对数据集在 $\boldsymbol{\Phi}(x)$ 空间上进行投影。

可知核函数主元分析方法只需要在原空间计算核函数，而不必去求 $\boldsymbol{\Phi}(x)$ 甚至不需要知道它的具体表达形式，这大大降低了计算的复杂度。值得注意的是原始数据和计算得到的矩阵 $\boldsymbol{K}$ 都应该做标准化处理，文献[11]对此做了详细的论述。目前常用的核函数有：

高斯径向核：$K(x_i,x)=\exp\left(\frac{-\parallel x-x_i\parallel^{2}}{2\delta^{2}}\right)$

多项式核：$K(x_i,x)=(x\cdot x_i+1)^{d},d=1,2,\cdots,N$

感知器核：$K(x_i,x)=\tan(\beta x_i+b)$

3.2　CA 原理

文化算法的主要思想是从种群中获取问题求解的经验知识并将其用于指导以后的搜索。总体上包括三大元素：种群空间(Population space)、信念空间(Belief space)和沟通渠道(Communication channel)，其中沟通渠道又包括接受函数(Acceptance function)、更新函数(Update function)、影响函数(Influence function)。种群空间从微观角度模拟个体根据一定的行为准则进化的过程，而信念空间则从宏观角度模拟文化的形成、传递和比较等进化过程，两个空间根据一定的通讯协议相互联系。文化算法的基本伪代码如下[12]：

```
Begin
t=0;
Initialize Population POP(t);
Initialize Belief Space BLF(t);
Repeat
Evaluate Population POP(t);
Update(BLF(t),Accept(POP(t)));
Variation(POP(t),Influence(BLF(t)));
t=t+1;
Select(POP(t) from POP(t-1));
Until termination condition achieved
End
```

文化算法框架提供了一种多进化过程的计算模型，因此从计算模型的角度来看，任何一种符合文化算法要求的进化算法都可以嵌入文化算法框架中作为种群空间的一个进化过程。

3.3　CA-KPCA 学习算法

KPCA 由于存在选择各种核函数及对应参数问题，如何根据具体问题选择最优核函数及参数，以达到最佳的分类聚类效果，是一个没有解决的问题。文化算法与核主元分析的有机结合，基本实现了对于不同的具体问题，选择适合该问题的最优核函数，达到最优分类聚类效果。

CA-KPCA 算法步骤：

(1)t=0；

(2)初始化种群规模核初始值、信念空间和相关参数、允许迭代次数或适应值限等；

(3)初始化设置核函数的类型、目标函数等；

(4)用 KPCA 提取特征,并将特征样本进行分类计算分类正确率即目标函数值;

(5)更新信念空间;

(6)根据父辈个体的适应值和信息空间的知识通过影响函数产生子代;

(7)$t=t+1$;

(8)通过锦标赛选择法从群体空间选出优秀个体;

(9)转到步骤(4),直至满足终止条件,即适应值误差达到设定的适应值误差限或迭代次数超过最大允许迭代次数,搜索停止,输出全局历史最优位置为所求核函数的最佳参数。

4 基于 CA-KPCA 的 FCM 算法

基于 CA-KPCA 的 FCM 算法通过两个阶段来实现。第一个阶段,将原始数据集通过 CA-KPCA 算法进行数据预处理,选出最优的核函数及参数,以及相应的主元。第二个阶段,将处理好的数据集利用 FCM 进行聚类,所要聚类的数目可由函数 $Sub(U;c)=\max_{l=1}^{c}\max_{h=1,h\neq l}^{c}(A_l,A_h)$的最小的 c 值来确定[13],其中:

$$C(A_l,A_h)=\frac{\sum_{i=1}^{n}(u_{il}^m d_{il}^2 \wedge u_{ih}^m d_{ih}^2)}{\sum_{i=1}^{n}(u_{il}^m d_{il}^2)} \tag{10}$$

模糊聚类的结果是对数据集进行模糊划分:$A=\{A_1,A_2,\cdots,A_c\}$(这里 A_l 表示样本属于 l 类的隶属函数),一个好的分类就应该使 A_l 与 A_h 尽可能分离,因此可以通过列举来比较式(10)的值,从而完成最佳的分类数目 c 的确定。

5 仿真实验

采用 UCI 提供的标准数据集 Breast cancer 和 wine 进行仿真实验。Breast cancer 数据集共有样本 683 个,每个样本含有10个属性,分成 2 类;wine 数据集共有样本 178 个,每个样本含有属性 13 个,分成 3 类。实验环境为 P4 2.66 GHz,内存 256 M,Matlab 7.0。

5.1 CA-KPCA 数据预处理

将 Breast cancer 和 wine 数据集用于建立核主元模型,并将其投影后分类,以分类错误率作为评价标准。选择高斯径向核为核函数,文化算法相关参数设置为种群规模 50、接受函数中%p=0.2、h=1 或 2,信念细胞分为 A、B、C 三类。结果如表 1 所示。

表 1 CA-KPCA 的结果

数据集	主元	运算时间/min	σ
Breast cancer	5	6.653	19.21
wine	6	7.012	19.08

5.2 FCM 聚类

将 CA-KPCA 预处理后的 Breast cancer 和 wine 数据集用 FCM 聚类,与传统的 FCM 算法进行比较,结果显示在表 2 中。

表 2 不同算法的比较

数据集	算法	精确度/(%)	训练时间/s
Breast cancer	FCM	95.75	0.892 102
	CA-KPCA-FCM	95.90	0.405 422
wine	FCM	48.31	0.825 818
	CA-KPCA-FCM	68.54	0.788 112

6 结论

KPCA 方法能有效地对数据进行特征提取,但核方法本身受到核参数选择的影响。通过文化算法对 KPCA 进行核函数优化选择,能提高特征提取的能力,降低分类错误率,再结合FCM 算法,使最终的 FCM 算法不论在速度上还是在精确度上都有了改善。

参考文献:

[1] Bezdek J C,Ehrlich R,Full W.FCM:The fuzzy C-means clustering algorithm[J].Computers and Geoscience,1984,10(2/3):191-203.

[2] Scholkope B,Alexander J S.Kernel principal component analysis [EB/OL].(1998-07-02)[2007-01-25].http://sml.nicta.com.an/~smola/papers/SchSmoMul99.pdf.

[3] Harandi M T,Ahmadabadia M N,Ara-abi B N.Optimal local basis:A reinforcement learning approach for face recognition[J].International Journal of Computer Vision,2008(8):161-165.

[4] 郭辉,王玲,刘贺平.基于核主成分分析与最小二乘支持向量机结合处理时间序列预测问题[J].北京科技大学学报,2006,28(3):303-306.

[5] 杨希,钱峰,张兵.基于核函数主元分析的 SVM 建模方法及应用[J].华东理工大学学报,2007,33(2):259-262.

[6] 黄海燕,柳桂国,顾幸生.基于文化算法的 KPCA 特征提取方法[J].华东理工大学学报,2008,34(2):256-260.

[7] Pineyro J,Klempnow A,Lescano V.Effectiveness of new spectral tools in the anomaly detection of rolling element bearings[J].Journal of Alloya and Compounds,2000,3(10):276-279.

[8] Deter W T,Peng Y H,Richard Y.Wavelet analysis and envelope detection for rolling element earning fault diagnosis:Their effectiveness and flexibilities[J].ASME J Vibr Acoust,2001,123:303-310.

[9] Yuan X H,Yuan Y B.Application of culture algorithm to generation scheduling of hydrothermal systems[J].Energy Conversion and Management,2006,47:2192-2201.

[10] 高丽丽,刘弘,李同喜.基于模式学习的文化遗传算法的研究[J].计算机工程与应用,2007,43(22):38-40.

[11] Scholkopf B,Smola A,Muller K R.Nonlinear component analysis as a kernel eigenvalue problem[J].Neural Computation,1998,10(5):1299-1319.

[12] Reynolds R G,Zhu Shinin.Knowledge-based function optimization using fuzzy cultural algorithms with evolutionary programming[J].IEEE Transactions on Systems,Man,and Cybernetics,Part B:Cybernetics,2001,31(1):1-18.

[13] 范九伦,吴成茂.FCM 算法中隶属度的新解释及其应用[J].电子学报,2004,32(2):350-352.