离群模糊核聚类算法*

沈红斌[1,2], 王士同[1+], 吴小俊[2,3]

[1](江南大学 信息学院,江苏 无锡 214036)

[2](华东船舶工业学院 计算机系,江苏 镇江 212003)

[3](中国科学院 沈阳自动化研究所 机器人学重点实验室,辽宁 沈阳 110015)

Fuzzy Kernel Clustering with Outliers

SHEN Hong-Bin[1,2], WANG Shi-Tong[1+], WU Xiao-Jun[2,3]

[1](School of Information, Southern Yangtse University, Wuxi 214036, China)

[2](Department of Computer, EastChina Shipbuilding Institute, Zhenjiang 212003, China)

[3](Robotics Laboratory, Shenyang Institute of Automation, The Chinese Academy of Sciences, Shenyang 110015, China)

+ Corresponding author: E-mail: wxwangst@yahoo.com.cn, http://www.pami.sjtu.edu.cn

Received 2003-08-11; Accepted 2003-10-08

Shen HB, Wang ST, Wu XJ. Fuzzy kernel clustering with outliers. *Journal of Software*, 2004,15(7): 1021~1029.

http://www.jos.org.cn/1000-9825/15/1021.htm

Abstract: Outliers are data values that lie away from the general clusters of other data values. It may be that an outlier implies the most important feature of a dataset. In this paper, a new fuzzy kernel clustering algorithm is presented to locate the critical areas that are often represented by only a few outliers. Through mercer kernel functions, the data in the original space are firstly mapped to a high-dimensional feature space. Then a modified objective function for fuzzy clustering is introduced in the feature space. An additional weighting factor is assigned to each vector in the feature space, and the weight value is updated using the iterative functions derived from the objective function. The final weight of a datum represents a kind of representativeness of the corresponding datum. With these weights, the experts can identify the outliers easily. The simulations demonstrate the feasibility of this method.

Key words: outlier; fuzzy; kernel function; feature space; clustering algorithm

摘 要: 一般说来,离群点是远离其他数据点的数据,但很可能包含着极其重要的信息.提出了一种新的离群模糊核聚类算法来发现样本集中的离群点.通过 Mercer 核把原来的数据空间映射到特征空间,并为特征空间的每

* Supported by the Jiangsu Key Laboratory of Computer Information Technology (江苏省计算机信息技术重点实验室开放课题); the National Key Laboratory for Novel Software Technology of Nanjing University (南京大学计算机软件新技术国家重点实验室开放课题); the Jiangsu Natural Science Foundation of China under Grant No.BK2003017 (江苏省自然科学基金)

作者简介: **沈红斌**(1979－),男,江苏句容人,博士,主要研究领域为模糊人工智能,数据挖掘;**王士同**(1964－)男,博士,教授,博士生导师,主要研究领域为人工智能,神经网络,模糊系统;**吴小俊**(1967－),男,博士,副教授,主要研究领域为人工智能,模式识别,神经模糊系统.

个向量分配一个动态权值,在经典的 FCM 模糊聚类算法的基础上得到了一个特征空间内的全新的聚类目标函数,通过对目标函数的优化,最终得到了各个数据的权值,根据权值的大小标识出样本集中的离群点.仿真实验的结果表明了该离群模糊核聚类算法的可行性和有效性.

关键词: 离群;模糊;核函数;特征空间;聚类算法

中图法分类号: TP18 **文献标识码**: A

离群点是样本集中一些远离其他数据点的数据,离群点可能是表明一些特殊的情况信息,对这种离群点要加以重点地研究和分析;也有可能是在收集数据的过程中造成错误从而导致了离群点的出现,对于这种离群点要剔除.由于离群点在整个数据集中只占很小的一部分,所以以前在处理数据的时候,往往会把这些离群点忽略掉,从而导致一些重要信息的丢失.近几年,对于离群点的发现越来越引起人们的关注.很多学者提出了不同的发现离群点的方法.总的说来,目前主要有两类发现离群点的方法:主观发现法和客观发现法.在主观方法中,用户直接把自己的知识融入定义参数的过程中,如"非常远"、"低密度"等,从这一点来说,主观发现方法对于不同的用户可能会导致不同的结果,并且其可扩展性也比较差[1],而且这种主观发现的方法只适用于一些小数据集,对于大数据集来说,将是一个非常耗时的过程.在客观发现的方法中,算法将根据数据点的分布情况,自动地发现一些离群点.文献[2]给出了一种基于图形方法的客观发现方法,这种方法通过构建一个 box plot 来描述数据样本的分布情况.实际上,已经有学者开始把模糊聚类算法应用于发现离群点的过程中,并取得了很好的结果[3],该算法是基于 FCM 模糊聚类算法,通过对每一个样本分配一个动态的权值,从而最终通过权值发现离群点,该算法对于线性可分的数据集能取得较好的效果.实际上,现实生活中常常会出现一些线性不可分但非线性可分的数据集,对于这种数据集,文献[3]的算法将不能很好地聚类并发现相应的离群点.我们知道,对于非线性可分的数据样本,可以在其特征空间内进行线性的聚类[4],很多非线性聚类技术就是采用了这种思想,如支撑向量机[5]、非线性区分分析[6]等.为此,本文提出了一种新的离群模糊核聚类算法,通过 Mercer 核函数,我们首先把数据样本空间映射到特征空间,对特征空间的每一个向量分配一个动态的权值,并在 FCM 模糊聚类的基础上得到一个新的包括向量权值的目标函数,在特征空间内,通过对该新目标函数的优化,最终得到了各个数据点的权值,从而根据这个权值发现数据样本中的离群点.该算法对于非线性可分的数据样本集仍能够给出令人满意的聚类效果,并发现样本集中的离群点.并且,也从理论上证明了离群模糊核聚类算法的收敛性.仿真实验验证了该算法的有效性和可行性.

1 离群模糊核聚类算法

1.1 Mercer核

设 $\vec{x}_k \in R^N (k=1,2,...,K)$ 是观察空间内的一组样本集,用一个非线性连续函数 ϕ 对这组样本进行映射,就能得到高维空间 H 中的一组向量集,$\phi(\vec{x}_1)$,$\phi(\vec{x}_2)$,…,$\phi(\vec{x}_K)$,这样,特征空间中的点积就能用观察空间中的核来表示:

$$K(\vec{x}_i,\vec{x}_j)=(\phi(\vec{x}_i)\cdot\phi(\vec{x}_j)) \tag{1}$$

所有的这些样本就组成了一个核函数矩阵 $K_{ij}=K(\vec{x}_i,\vec{x}_j)$ [7],这是众多非线性分类技术的基础,如支撑向量机[4]、非线性区分分析[5]等.核函数具有以下两个特征:

- 对称性

$$K(\vec{x},\vec{y})=K(\vec{y},\vec{x}) \tag{2}$$

- 满足 Cauchy-Schwarz 不等式

$$K(\vec{x},\vec{y})^2 \le K(\vec{x},\vec{x})K(\vec{y},\vec{y}) \tag{3}$$

目前,对于核函数的选择还没有一个通用的标准,常用的核函数有以下几种:

- 多项式核函数

$$K(\vec{x},\vec{y})=(\vec{x}\cdot\vec{y}+1)^d \tag{4}$$

其中 d 是用户定义的整数.

- 高斯核函数

$$K(\vec{x},\vec{y})=\exp\left(\frac{-\|\vec{x}-\vec{y}\|^2}{2\sigma^2}\right) \tag{5}$$

其中σ为高斯函数的宽度.

- 二层神经网络 sigmoidal 核函数

$$K(\vec{x},\vec{y})=\tanh(-b(\vec{x}\cdot\vec{y})-c) \tag{6}$$

其中 b,c 是用户自定义的参数.

上述高斯核函数是最常用的核函数,因为高斯核函数所对应的特征空间是无穷维的,有限的样本在该特征空间中肯定是线性可分的.在本文中,将采用高斯核函数作为映射函数.

1.2 离群问题

对于离群问题的定义,很多学者对于不同问题提出了不同的定义,但其内在含义是一致的,即离群点是在收集数据的过程中由于误操作或者异常情况的出现,从而出现了一些不符合正常特征的数据点.对于这些离群点,如果是由于误操作而得到的,则要剔除掉,从而提高数据样本的质量;如果是由于出现了新规律而导致离群点的产生,则应该对这些点加以重点的分析.在以前的数据处理中,常常会把离群点当作小概率事件而不加以处理,从而造成了很大的损失.当前,对于发现离群点越来越引起人们的注意.图 1 中类似 1,2 两点的数据与其他数据样本的性质不同,本文所提出的离群模糊核聚类算法就是为了发现这些离群数据以供专家对其作进一步的分析.

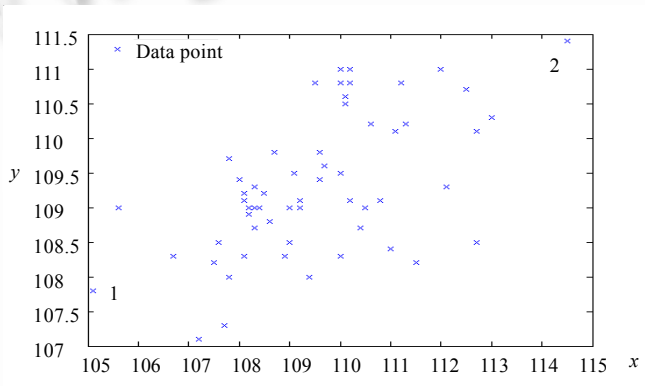

Fig.1 A sample dataset with outliers

图 1 含有离群点的数据样本

1.3 离群模糊核聚类算法

许多聚类的算法是基于平方误差和的方法[8,9].给定观察空间中的一个有限数据集合 $\vec{x}_k\in R^N(k=1,2,...,K)$,通常的目标函数定义成如下形式:

$$J(X,U,v)=\sum_{i=1}^{C}\sum_{k=1}^{K}{\mu_{ik}}^m(\vec{x}_k-\vec{v}_i)(\vec{x}_k-\vec{v}_i)^T \tag{7}$$

其中 C 是类别数,$\mu_{ik}\in[0,1]$ 是 $\vec{x}_k$ 属于第 i 类的隶属度,$\vec{x}_k$ 为第 k 个样本向量,$\vec{v}_i$ 表示第 i 类中心向量.为了更方便地发现离群点,文献[3]为每一个样本点分配了一个动态的权值,得到了如下目标函数:

$$J(X,U,v)=\sum_{i=1}^{C}\sum_{k=1}^{K}{\mu_{ik}}^m\frac{1}{{w_k}^q}(\vec{x}_k-\vec{v}_i)(\vec{x}_k-\vec{v}_i)^T \tag{8}$$

其中 w_k 是第 k 个样本的权值,且 $\sum_{k=1}^{K}w_k=w$,w 为用户定义的一个常量.

当用一个平滑而连续的非线性核函数 ϕ 把数据样本集映射到高维空间 H 时,

$$\phi:\quad R^N\to H\qquad \vec{x}\to\vec{X},$$

原数据空间中的数据样本的拓扑结构将保持不变[6],所以,式(8)目标函数在特征空间 H 中就可以写成如下形式:

$$J_H=\sum_{i=1}^{C}\sum_{k=1}^{K}{\mu_{ik}}^m\frac{1}{{w_k}^q}(\phi(\vec{x}_k)-\vec{m}_i^{\ \phi})(\phi(\vec{x}_k)-\vec{m}_i^{\ \phi})^T \tag{9}$$

其中 $\vec{m}_i^{\ \phi}$ 表示特征空间 H 中第 i 类的中心.

应该注意到,J_H 是由特征空间中的一系列的点积所构成的,正如前面所述,内积可以通过核函数来进行运算.通过一个特定的核函数,内积就定义了一个到特征空间的非线性映射 ϕ[4],所以特征空间中的平方和准则就可以根据一个对称的 $K\times K$ 的核矩阵的元素写出,式(9)所定义的目标函数就可以写成

$$J_H=\sum_{i=1}^{C}\sum_{k=1}^{K}{\mu_{ik}}^m\frac{1}{{w_k}^q}Q_{ik} \tag{10}$$

其中,

$$Q_{ik}=K_{kk}-\frac{2}{N_i}\sum_{j=1}^{K}\mu_{ij}K_{kj}+\frac{1}{{N_i}^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\mu_{ij}\mu_{il}K_{jl} \tag{11}$$

其中,$N_i=\sum_{k=1}^{K}\mu_{ik}$,$K_{ij}=k(\vec{x}_i,\vec{x}_j)$ 是所用的核函数,Q_{ik} 表示特征空间中第 k 个向量到第 i 类中心的距离.式(10)将作为离群模糊核聚类算法的优化目标函数.

众多研究成果已经表明,离群信息往往是数据样本集中最重要的特征[3],所以离群模糊核聚类算法的目标是对普通数据样本分配一个小权值 w_k ($\frac{1}{{w_k}^q}$ 就大),而一般地,离群点将远离任何一个类中心,从而分配一个大权值 w_k ($\frac{1}{{w_k}^q}$ 就小)给离群点.参数 q 在聚类过程中起到重要的作用:当 q 足够大时,每个数据样本的权重值将趋近相等 $\frac{w}{K}$,也就是说,权重对于所有的样本都有相同的影响;当 $q\to 0$ 时,权重影响将达到最大.

下面导出 w_k 的迭代公式,考虑限制条件 $\sum_{k=1}^{K}w_k=w$,由 Lagrange 函数可以得到下面的无限制的方程:

$$J_H=\sum_{i=1}^{C}\sum_{k=1}^{K}{\mu_{ik}}^m\frac{1}{{w_k}^q}Q_{ik}+\lambda(\sum_{k=1}^{K}w_k-w) \tag{12}$$

式(12)对 w_k 求偏微分可得

$$\frac{\partial J_H}{\partial w_k}=-q\cdot\frac{1}{{w_k}^{q+1}}\cdot\sum_{i=1}^{C}{\mu_{ik}}^m\cdot Q_{ik}+\lambda \tag{13}$$

令 $\frac{\partial J_H}{\partial w_k}=0$,可求得

$$\lambda=q\cdot\frac{1}{{w_k}^{q+1}}\cdot\sum_{i=1}^{C}{\mu_{ik}}^m\cdot Q_{ik} \tag{14}$$

由上式可求得

$$w_k=\left(\frac{q\cdot\sum_{i=1}^{C}{\mu_{ik}}^m\cdot Q_{ik}}{\lambda}\right)^{\frac{1}{q+1}} \tag{15}$$

根据 $\sum_{k=1}^{K}w_k=w$,由式(15)得

$$w=\sum_{k=1}^{K}\left(\frac{q\cdot\sum_{i=1}^{C}{\mu_{ik}}^m\cdot Q_{ik}}{\lambda}\right)^{\frac{1}{q+1}} \tag{16}$$

所以,

$$\lambda^{\frac{1}{q+1}}=\sum_{k=1}^{K}(q\cdot\sum_{i=1}^{C}{\mu_{ik}}^{m}\cdot Q_{ik})^{\frac{1}{q+1}}\cdot\frac{1}{w} \tag{17}$$

$$\lambda=\left(\sum_{k=1}^{K}\left(q\cdot\sum_{i=1}^{C}{\mu_{ik}}^{m}\cdot Q_{ik}\right)^{\frac{1}{q+1}}\cdot\frac{1}{w}\right)^{q+1} \tag{18}$$

由式(14)、式(18)可得

$$w_k=\frac{\left(q\cdot\sum_{i=1}^{C}{\mu_{ik}}^{m}\cdot Q_{ik}\right)^{\frac{1}{q+1}}}{\sum_{k=1}^{K}\left(q\cdot\sum_{i=1}^{C}{\mu_{ik}}^{m}\cdot Q_{ik}\right)^{\frac{1}{q+1}}}\cdot w=\frac{\left(\sum_{i=1}^{C}{\mu_{ik}}^{m}\cdot Q_{ik}\right)^{\frac{1}{q+1}}}{\sum_{k=1}^{K}\left(\sum_{i=1}^{C}{\mu_{ik}}^{m}\cdot Q_{ik}\right)^{\frac{1}{q+1}}}\cdot w \tag{19}$$

这样,就导出了 w_k 的迭代公式.在对目标函数的优化过程中,很多学者提出了多种优化的方法,比如,文献[8,10]提出了推测优化算法等.仿照著名的 FCM 模糊聚类算法,根据上述 Q 度量,很容易推导出下面的隶属度迭代公式:

$$\mu_{ik}=\frac{1}{\sum_{j=1}^{C}\left(\frac{Q_{ik}}{Q_{jk}}\right)^{\frac{1}{m-1}}} \tag{20}$$

在考虑类中心的迭代公式时,须考虑权值的关系:

$$m_i=\frac{\sum_{k=1}^{K}\langle\mu_{ik}\rangle\bar{x}_k}{\sum_{j=1}^{K}\langle\mu_{ij}\rangle} \tag{21}$$

其中,$\langle\mu_{ik}\rangle=\frac{{\mu_{ik}}^{m}}{{w_k}^{q}}$.

离群模糊核聚类算法通过在高维空间进行模糊聚类,得到了相应数据样本的权值,从而发现数据样本集中的离群信息,其复杂度与著名的 FCM 算法相同,也可以仿照 FCM 算法的相关证明而得到.该离群模糊核聚类算法可以总结如下:

离群模糊核聚类算法 FKCO(fuzzy kernel clustering with outliers).

(1) 初始化算法参数 C,q,m,ε,其中 C 为聚类数目,q 为权重指数,m 为模糊因子,ε为一个很小的正数,设算法迭代记数 t=1.

(2) 初始化样本集中样本隶属度参数.

(3) 根据式(11)计算向量样本到类中心距离 $Q_{ik}^{(t)}$.

(4) 根据 $Q_{ik}^{(t)}$,重新计算每一样本向量的隶属度 $\mu_{ik}^{(t)}$.

(5) 根据式(19)计算得到每个样本向量权重系数 $w_k^{(t)}$.

(6) 如果 $\left|(J_H(t)-J_H(t-1))\right|>\varepsilon$,$t\leftarrow t+1$,转(3).

否则,算法终止.

下面给出离群模糊核聚类算法的收敛性证明.

定理 1. 在离群模糊核聚类算法中,$\mu_{ij}(i=1,2,...,C,j=1,2,...,K)$ 和 $w_j(j=1,2,...,K)$ 是 J_H 局部最优的必要条件是 $\mu_{ij}=\frac{1}{\sum_{k=1}^{C}\left(\frac{Q_{ij}}{Q_{kj}}\right)^{\frac{1}{m-1}}}$ 满足 $\sum_{i=1}^{C}\mu_{ij}=1$ 且 w_j 由式(19)得到,满足 $\sum_{j=1}^{K}w_j=w$ 限制条件.

证明:首先假设 w_j 固定,问题就变成求 J_H 对 μ_{ij} 的最小值且满足 $\sum_{i=1}^{C}\mu_{ij}=1$ 限制条件由 Lagrange 函数可得:

$$L(W,\lambda)=J_H-\sum_{j=1}^{K}\lambda_j\left(\sum_{i=1}^{C}\mu_{ij}-1\right) \tag{22}$$

求得偏微分方程:

$$\frac{\partial L(W,\lambda)}{\partial \mu_{ij}}=m{\mu_{ij}}^{m-1}\frac{1}{{w_j}^q}Q_{ij}-\lambda_j=0 \tag{23}$$

$$\frac{\partial L(W,\lambda)}{\partial \lambda_j}=\sum_{i=1}^{C}\mu_{ij}-1=0 \tag{24}$$

则由式(23)可解得

$$\mu_{ij}=\left[\frac{\lambda_j{w_j}^q}{mQ_{ij}}\right]^{\frac{1}{m-1}} \tag{25}$$

把式(25)代入式(24)得

$$\left(\frac{\lambda_j{w_j}^q}{m}\right)^{\frac{1}{m-1}}=\frac{1}{\left(\sum_{k=1}^{C}\frac{1}{Q_{kj}}\right)^{\frac{1}{m-1}}} \tag{26}$$

把式(26)代入式(25)得到隶属度迭代函数式(20).

要证明 w_j 由式(19)得到,满足 $\sum_{j=1}^{K}w_j=w$ 限制条件,其过程类似于 w_j 迭代公式的推导,这里从略. □

定理 2. 设 $\phi(U)=J_H$,其中 $U=[\mu_{ij}]_{C\times K}$, $w_j(j=1,2,...,K)$ 固定且对所有的 $1\le i\le C,1\le j\le K$ 有 $Q_{ij}\ne 0$,那么,U 是 $\phi(U)$ 的一个局部最优,当且仅当 w_j $(j=1,2,...,K)$ 由式(19)算出.

证明:必要性已由定理 1 证明得到.要证明其充分性,考虑由式(20)得到的 Lagrange $\phi(U)$ 的 Hessian 矩阵 $H(\phi)$.由式(22)可得

$$h_{st,ij}(U)=\frac{\partial}{\partial\mu_{st}}\left[\frac{\partial\phi(U)}{\partial\mu_{ij}}\right]=\begin{cases}m(m-1){\mu_{ij}}^{m-2}\dfrac{1}{{w_j}^q}Q_{ij}, & \text{if } s=i,t=k\\ 0, & \text{otherwise}\end{cases} \tag{27}$$

其中 μ_{st} 由式(20)得到.从式(27)可以看出 $H(U)=[h_{st,ij}(U)]$ 为一对角阵.由于对所有的 $1\le i\le C,1\le j\le K$, 都有 $m>1, Q_{ij}>0, w_j>0$,所以上述 Hessian 矩阵 $H(U)$ 为一个正定阵.所以式(20)为最小化 $\phi(U)$ 的充分条件. □

定理 3. 设 $\phi(W)=J_H$, $U=[\mu_{ij}]_{C\times K}$ 固定,对所有 $1\le i\le C$, $1\le j\le K$ 都有 $Q_{ij}\ne 0$, $m>1$,那么 $w_j(j=1,2,...,K)$ 是 $\phi(W)$ 的局部最优,当且仅当 $w_j(j=1,2,...,K)$ 由式(19)所计算出.

证明:必要性由前面的定理已经证明出.要证明其充分性,由式(12)可得:

$$\frac{\partial}{\partial w_i}\left[\frac{\partial\phi(w)}{\partial w_j}\right]=\begin{cases}q(q+1)\sum_{k=1}^{C}{\mu_{kj}}^m\dfrac{1}{{w_j}^{q+2}}Q_{ij}>0, & \text{if } \quad i=j\\ 0, & \text{otherwise}\end{cases} \tag{28}$$

也就是说,其 Hessian 矩阵为一个正定阵,所以式(19)是优化 $\phi(W)$ 的充分条件. □

由定理 2、定理 3 能够证明得到

$$J_H(U^{t+1},W^{t+1})\le J_H(U^t,W^t) \tag{29}$$

也就是说,J_H 是 t 的递减函数,所以,离群模糊核聚类算法将最终收敛.

2 仿真实验

为了测试本文提出的模糊核聚类算法的性能,分别用线性可分、线性不可分以及高维的数据样本集进行测试.实验表明,模糊核聚类算法对这 3 种数据样本集都能取得令人满意的结果.

实验 1. 首先,我们把离群模糊核聚类算法应用于线性可分的样本集.图 2 显示了该数据样本集.我们对该样本集进行了 3 种算法的实验:FCM(如图 3 所示)、文献[3]中的算法(如图 4 所示)以及离群模糊核聚类算法(如图 5 所示).设 m=2,q=1,w=200.

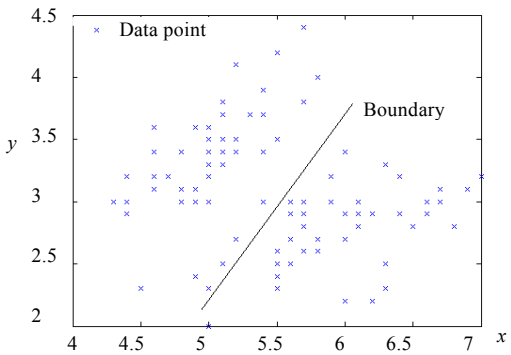

Fig.2 A testing linear separable dataset

图 2 一个线性可分的数据样本集

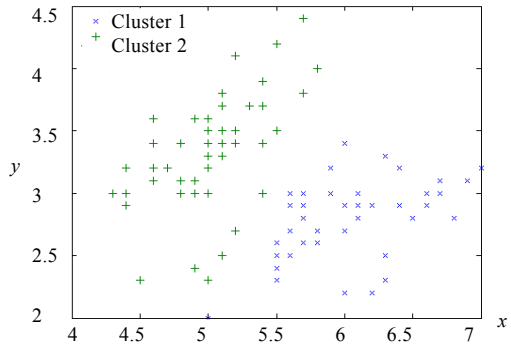

Fig.3 The clustering result of FCM

图 3 FCM 聚类的效果

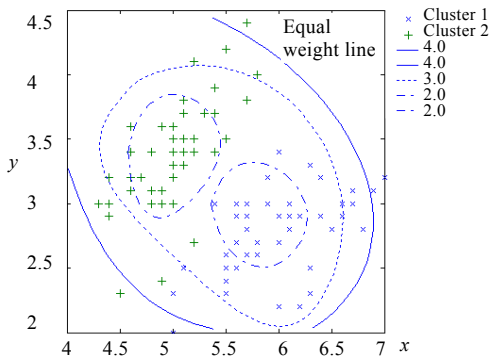

Fig.4 Result of the algorithm in Ref.[3] identifying outliers on the dataset with the weights

图 4 文献[3]种算法对该数据样本集的聚类效果以及根据等权重线标识离群点的结果图

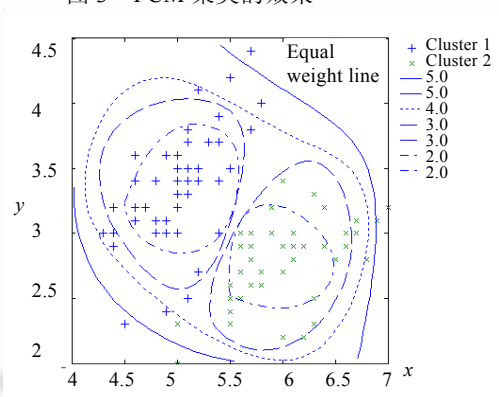

Fig.5 Result of FKCO algorithm identifying the outlier information

图 5 FKCO 算法对该数据样本集的聚类效果以及根据等权重线发现离群点的结果图

从图 4、图 5 可以看出,当算法结束时,根据最终得到的权值,可以作出等权重线,在一定的等权重线之外的数据点就被认为是该样本集在该权重下的离群点,权重的选择由专家给出.对于线性可分的数据样本集,离群模糊核聚类算法能对数据样本集取得令人满意的聚类效果,并在发现离群点的过程中,取得了与文献[3]相似的结果,但从结果看出,本文所提离群模糊核聚类算法其权值变化幅度较小,也就是说,对待离群点更谨慎些,而这一点更符合人们对待离群点的正常思维方式.

实验 2. 在这个实验中,一个线性不可分的数据样本集被用来测试离群模糊核聚类算法的性能.图 6 给出了该数据样本集,图 7 显示了 FCM 的聚类效果,文献[3]中的算法的聚类效果如图 8 所示.从图 7、图 8 可以看出,由于 FCM 算法和文献[3]中算法均是在观察空间中聚类,所以这两种算法对于此种样本集将不能给出令人满意的聚类.文献[11]中的硬核聚类*算法*的聚类效果如图 9、图 10 给出了离群模糊核聚类算法的分类效果以及根据等权重线标识离群点的结果,由于这两种算法均为在特征空间中进行聚类,所以虽然算法不一样,但都能取得较

好的聚类效果.另外,由于本文提出的算法加入了权值的概念,所以能够更容易地发现样本集中的离群点.设参数 m=2,q=1,w=200.

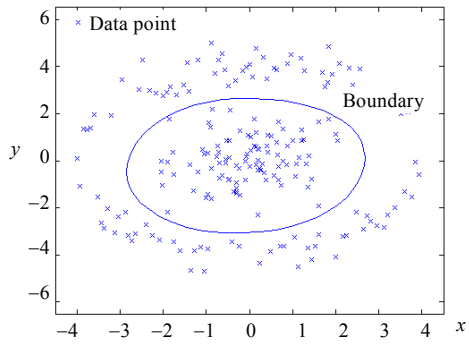

Fig.6 Linear inseparable but nonlinear separable dataset

图 6 实验用线性不可分数据集

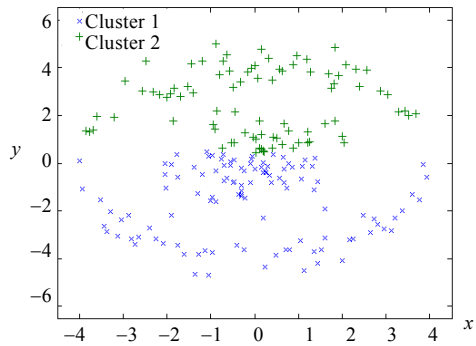

Fig.7 Clustering performance of FCM

图 7 FCM 的聚类效果

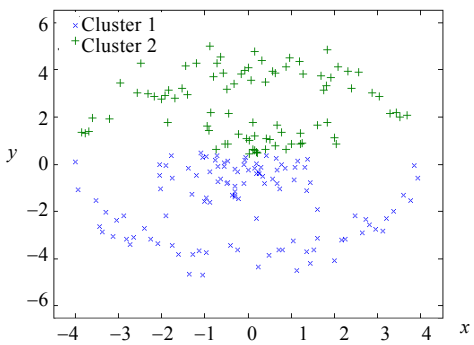

Fig.8 Clustering performance of the algorithm in Ref.[3]

图 8 文献[3]算法的聚类效果图

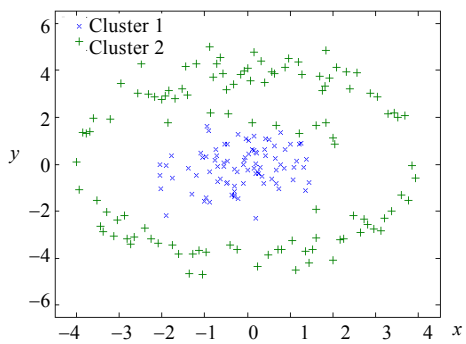

Fig.9 Clustering performance of the algorithm in Ref.[11]

图 9 核聚类算法[11]的聚类效果

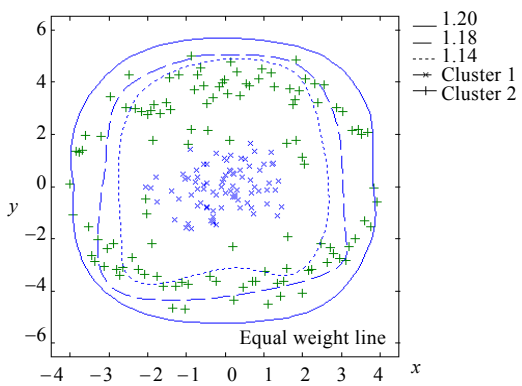

Fig.10 Performance of the clustering and identifying the outliers of FKCO

图 10 离群模糊核聚类算法的聚类以及发现离群点的结果

从上面的结果可以看出,对于线性不可分的数据样本集,传统的聚类算法如 FCM 以及其变形算法[3]将要失败,当把数据空间用核函数映射到其特征空间以后,就能取得较好的聚类效果.根据最终得到的权重,离群模糊

核聚类算法就能很容易地发现离群信息.

实验 3. 这个实验将把著名的 IRIS 数据样本集作为测试样本测试本文所提算法处理高维数据样本集的能力.表 1 对比了本文提出的 FKCO 算法以及文献[3]中算法分别对于 IRIS 数据样本集的聚类效果.从结果可以看出,当映射到高维空间以后,FKCO 对于高维的 IRIS 数据样本集聚类结果要优于 FCM 以及文献[3]中的算法.

Table 1 Comparison of clustering performances on IRIS dataset

表 1 IRIS 数据样本集聚类结果对比

Algorithms	Numbers of data wrongly clustered
FCM	15
FKCO	14
The algorithm in Ref.[3]	16

众所周知,研究已经发现,在IRIS数据样本集中没有明显的离群信息,也就是说权值的变化范围越小越能体现这一特性.为了证明这一点,表 2 分别给出了 FKCO 算法以及文献[3]中算法权重变化范围 $V_w = \max_{i,j}(w_i - w_j)$,由表 2 可以看出,FKCO 算法的权重变化范围要小于文献[3]中算法权值变化范围.所以这一点也从另外一个侧面说明了当映射到高维空间以后,算法具有更强的分辨效率.

Table 2 Weight's change scopes between two algorithms

表 2 两种算法权值变化区域 V_w 比较

Algorithms	Weight's change scope V_w
FKCO	2.979 6
The algorithm in Ref.[3]	3.365 4

3 结 论

本文提出了一种新的离群模糊核聚类算法.该算法首先用核函数把原数据空间映射到特征空间,通过为特征空间中的向量分配一个动态的权值,最终在得到好的聚类的同时,能借助权值发现样本集中的离群点;特别是对于一些线性不可分的样本集,在运用传统算法失败的情况下,离群模糊核聚类算法仍然能在取得良好的聚类效果的同时发现其离群点.同时,我们也证明了离群模糊核聚类算法的收敛性.仿真实验表明了离群模糊核聚类算法的有效性.

References:

[1] Last M, Kandel A. Automated perceptions in data mining. In: Proc. of the 8th Int'l Conf. on Fuzzy System. Seoul, 1999. 190~197.

[2] Mendenhall W, Reinmuth JE, Beaver RJ. Statistics for Management and Economics. 6th ed., Belmont: Duxbury Press, 1993.

[3] Keller A. Fuzzy clustering with outliers. In: Proc. of the NAFIPS00. 2000. 143~147.

[4] Girolami M. Mercer kernel-based clustering in feature space. IEEE Trans. on Neural Networks, 2002,13(3):780~784.

[5] Vapnik VN. The Nature of Statistical Learning Theory. 2nd ed., New York: John Wiley and Sons, 1998.

[6] Roth V, Steinhage V. Nonlinear discriminant analysis using kernel functions. In: Solla SA, Leen TK, Muller K-R, ed. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 1999. 568~574.

[7] Schölkopf B, Mika S, Burges CJC, Knirsch P, Müller K-R, Rätsch G, Smola AJ. Input space versus feature space in kernel-based methods. IEEE Trans. on Neural Networks, 1999,10(5):1000~1017.

[8] Buhmann JM. Data clustering and data visualization. In: Jordan MI, ed. Learning in Graphical Models. Boston: Kluwer, 1998.

[9] Höppner F, Klawonn F, Eklund P. Learning indistinguishability from data. Soft Computing, 2002,6(1):6~13.

[10] Roberts SJ, Everson R, Rezek I. Maximum certainty data partitioning. Pattern Recognition, 2000,33(5):833~839.

[11] Zhang L, Zhou WD, Jiao LC. Kernel clustering algorithm. Chinese Journal of Computers, 2002,25(6):587~590 (in Chinese with English abstract).

附中文参考文献:

[11] 张莉,周伟达,焦李成.核聚类算法.计算机学报,2002,25(6):587~590.