基于核的PP主成分分析及其在离群聚类中的应用*)

徐雪松 张 谱 宋东明 张 宏 刘凤玉
(南京理工大学计算机科学与技术学院 南京 210094)

摘 要 为了提高高维数据集合离群数据挖掘效率,在分析了传统的离群数据挖掘算法优点和缺点的基础上,提出了一种离群聚类算法,该算法将核方法与PP主成分变换结合于离群聚类算法中,采用基于核的PP主成分变换进行数据维数消减。通过该数据变换矩阵得到相应的非线性向量,并为每个向量分配一个动态权值,在优化经典的FCM模糊聚类的目标优化迭代函数基础上,最终得到各个数据的权值,根据权值的大小标识出数据集中的离群点,理论上证明了该算法的收敛性,仿真实验的结果表明了该方法能够有效地发现高维数据集中的离群点。

关键词 核方法,投影寻踪,主成分,模糊聚类,离群数据

The PP Principal Component Based on Kernel and its Application in Clustering with Outliers

XU Xue-Song ZHANG Xu SONG Dong-Ming ZHANG Hong LIU Feng-Yu
(Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094)

Abstract The data dimension reduction is the main method that can enhance the outliers mining efficiency based on higher-dimension data set. A novel clustering with outlier algorithm that is a combination of the kernel method and PP principal component is proposed after analyzing the advantages and disadvantages of the classical outlier mining algorithm in the paper. In this paper, we introduce data transformation of PP principal component based on kernel to reduce data dimension. Through the data transformation matrix, we can obtain nonlinear data dimension and add an additional weighting factor for each vector. On the basis of modifying iterative functions derived from objective function for fuzzy clustering, the final weight value of a datum represents a kind of representativeness of the corresponding datum. With these weight values, the experts can identify the outliers easily. Thetheoretical analysis indicate that the algorithm is converged finally. Simulation results illustrate that this algorithm is very efficient.

Keywords Kernel method, Projection pursuit, Principal component, Fuzzy clustering, Outliers

1 引言

离群数据(outlier)就是明显偏离其它数据,不满足数据的一般模式或行为,与存在的其它数据不一致的数据[1]。离群数据挖掘就是从大量复杂的数据集中发现存在于小部分异常数据中的新颖的、与常规数据模式显著不同的新的数据模式。目前已经出现了一些高效的离群数据挖掘算法。它们可分为基于聚类的、统计的、距离的、深度的以及基于密度的方法等五种类型[2~6]。与统计学中的离群值稍有不同,统计学中的离群值往往指的是一维的数据,而我们要研究的数据是高维的。研究表明,高维空间下数据的特征完全不同于低维数据,因此高维空间中的离群点发现技术必然不同于传统的离群点发现方法。高维空间数据分布稀疏,这样高维数据的稀疏分布使数据之间的距离尺度及以此为基础的区域密度不再具有直观意义。从一个数据点来看,其他点到它的距离落在一个很小的区间内,很难给出一个合适的近似度阈值来确定哪些点与之相似[7]。另外,在实际使用时数据维数往往很高,而对高维数据的估计需要的样本个数与维数构成指数增长的关系,这在机器学习中称作著名的"维数灾难"(Curse of Dimensionality)。如在图像识别时,我们会使用一幅图像的所有像素点;在互联网信息检索时,一般对一个网页上常用词组进行词频统计,得到对此网页的描述。直接在高维数据上寻找数据间的离群点往往会带来严重的计算问题。为了解决这一问题,一些新的研究开始将高维空间的数据投影到子空间以后再进行离群点检测。例如,美国IBM公司的研究员Aggarwal[8]等采用演化计算方式寻找所有投影到子空间稀疏的小方格,将其中的数据作为离群点。国内复旦大学的一些学者在离群点发现技术上做了大量工作,如提出基于估计的高效子空间局部离群点发现等[9],这个算法能够在预先去除大量不可能成为离群点的对象前提下找到在所有子空间中的所有局部离群点,可以降低计算量。文[10]不是平等地看待各个维而是根据某些标准选择一些维,在这些维组成的子空间中寻找离群点。现有的研究表明,将数据投影到子空间再进行数据挖掘是可行的,但这带来了另一个问题——由于在实际使用中,样本点是非常稀少的,而随着数据维数的增加,对维度进行组合得到的子空间个数呈指数级增长。对此我们不可能采用穷举法,对每一个可能的子空间进行投影,再从中选择效果最好的子空间,因为这样的计算代价太大。另

*)基金项目:国家自然科学基金资助项目(60273035);国防科工委基础应用项目(K1704060511)。徐雪松 博士生,主要研究方向:离群数据发现技术,信息安全;张 谱 博士生,主要研究方向:分布式虚拟现实;宋东明 博士生,主要研究方向:系统仿真;张 宏 教授,博士生导师,主要研究方向:信息安全;刘凤玉 教授,博士生导师,主要研究方向:人工智能与信息安全。

外，大量的数据分析问题本质上是非线性的，甚至是高度非线性的，直接投影到子空间进行计算将不能很好地发现离群点。投影寻踪(Projection Pursuit)是最常用的特征提取方法，被广泛应用在各种领域，如模式识别、图像处理、性能保持、故障诊断等。它通过对原始数据加工处理，简化问题处理的难度并提高数据信息分析效率。我们知道核函数在支持向量机(SVM)中起着很重要的作用，它是解决非线性问题及克服维数灾难问题的关键。它的基本原理是将低维的输入空间数据映射到特征空间，使用一个非线性变换将一个线性不可分的空间映射到一个高维线性可分的空间。支持向量机算法的技巧在于不直接计算复杂的非线性变换，而是计算非线性变换的点积，即核函数，从而大大简化了计算。由此，本文提出了一种思路——将基于核的 PP 主成分数据变换应用于离群聚类算法中，将 PP 主成分处理所得的低维数据向量与核函数有机融合，形成非线性数据变换，在特征空间中对每一个向量分配一个动态的权值，通过对经典的模糊聚类改造基础上得到一个新的包括向量权值的目标优化迭代函数，在特征空间对该新的目标优化迭代函数进行优化，并使算法在离群聚类过程中对权值进行自适应调整，最终发现数据样本点中的离群点。本文也从理论上证明了提出的离群聚类算法的收敛性，仿真实验证实了该算法比传统的离群数据发现算法更加快捷，效果更好。

本文第 2 节介绍基于核的 PP 主成分分析；第 3 节介绍提出的离群聚类算法；第 4 节给出算法的收敛性分析；第 5 节是实验及其结果分析；最后是结束语。

2 基于核的 PP 主成分分析

2.1 PP 主成分分析与核方法

PP 主成分的基本思想是将高维数据投影到低维(1～3 维)子空间上，寻找能反映原来高维数据的结构或特征的投影，以达到研究、分析高维数据的目的[11]。采用 PP 主成分方法处理高维数据的优点主要体现在以下 3 个方面[12]：

(1)PP 主成分方法没有受到一般降维方法要求正态分布的假设的约束，而且事实上，自然科学中有许多数据是不符合正态分布的，或者人们对数据的分布没有足够的先验知识。

(2)克服了由于“维数灾难(Curse of Dimensionality)”带来的问题，同时也增加了数据的可视性。

(3)可以排除与数据结构和特征无关的或关系甚小的变量干扰。

它的缺点主要是计算量大，处理非线性能力弱。

核方法作为一种由线性到非线性之间的桥梁，起源于上世纪初叶[13]，核方法的基础是选择一个对称、连续且满足 Mercer 条件的函数 $K(x,y)$。在非线性空间中，只需考虑高维特征空间的点积运算 $\Phi(x)\cdot\Phi(y)=K(x,y)$，不必明确知道 $\Phi(x)$是什么函数。本文选用最常用的高斯核函数，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。

$$K(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{1}$$

2.2 基于核的 PP 主成分分析

假设输入空间数据集 $x=(x_1,x_2,\cdots,x_n)$，$x_k\in R^d$，其样本协方差矩阵为

$$C=\frac{1}{n}\sum_{j=1}^{n}x_jx_j^T \tag{2}$$

又记 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_k$ 为 C 的特征值，则所求的向量 w_1 就是 C 相应于 λ_1 的特征向量，这就得到了 $x_1,x_2,\cdots,x_n$ 的第一 PP 主成分 $w_1^Tx_1,w_1^Tx_2,\cdots,w_1^Tx_n$，其标准差为 λ_1，继续求垂直于 w_1 的特征向量 w_2，使

$$\sigma(w_2^Tx_1,w_2^Tx_2,\cdots,w_2^Tx_n)=\max_{w_1\perp w_2}\sigma((w^Tx_1,w^Tx_2,\cdots,w^Tx_n) \tag{3}$$

得到的 w_2 是相应于 λ_2 的特征向量和第二 PP 主成分 $w_2^Tx_1$，$w_2^Tx_2,\cdots,w_2^Tx_n$，其标准差是 λ_2，如此反复，直到某个 PP 主成分的标准差接近于零为止。若 λ_{k+1} 的值小到可以忽略，就意味着数据主要分布在由 $w_1,w_2,\cdots,w_k$ 所张成的 k 维子空间里，从而降低了原数据的维数。PP 主成分的目标是找到向量 w 使投影 $w\cdot x$ 具有最大的偏差，相当于 1 类 Fisher 判别函数问题，即

$$\max_{w}\sum_{i=1}^{n}(w,x_i)^2 \tag{4}$$

为使 $\|w\|$ 尽量小，优化问题的最小化目标函数[14]为

$$R(w,e)=\frac{1}{2}w\cdot w-\frac{\gamma}{2}\sum_{i=1}^{n}e_i^2 \tag{5}$$

其条件为 $e_i=w\cdot x_i,i=1,\cdots,n$。引入拉格朗日函数求解后得到 $\frac{1}{\gamma}a_i-\sum_{j=1}^{n}a_jx_j\cdot x_i=0$ (6)

定义 $\lambda=1/\gamma$，得到对偶的对称特征值问题

$$\begin{bmatrix}x_1\cdot x_1 & \cdots & x_1\cdot x_n\\ & \vdots & \\ x_n\cdot x_1 & \cdots & x_n\cdot x_n\end{bmatrix}\begin{bmatrix}a_1\\ \vdots\\ a_n\end{bmatrix}=\lambda\begin{bmatrix}a_1\\ \vdots\\ a_n\end{bmatrix} \tag{7}$$

为得到最大的偏差，应选择最大特征值对应的特征向量，投影量为

$$P(x)=w\cdot x=\sum_{i=1}^{n}a_ix_i\cdot x \tag{8}$$

引入非线性变换 $\Phi(x)$[14]，将投影放在高维特征空间进行，则对偶问题变成

$$Ca=\lambda a \tag{9}$$

式中，$C_{ij}=\Phi(x_i)\cdot\Phi(x_j)=K(x_i,x_j)$，投影量变成

$$P(x)=w\cdot\Phi(x)=\sum_{i=1}^{n}a_iK(x_i\cdot x) \tag{10}$$

因此，原始输入数据集 x 经过基于核的 PP 主成分数据变换在特征空间形成了一系列的点积。下面将该变换所得的投影函数 $P(x)$引入本文所提出的离群聚类算法中，其不仅具有优秀的主元提取性能，尤其适合于处理高维非线性问题。

3 离群聚类算法

在众多的模糊聚类算法中，应用最为广泛而且较成功的是 1974 年由 Dunn 提出并由 Bezdek 加以推广的模糊 C-均值(fuzzy C-means)算法。其目标优化迭代函数定义如下：

$$J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^md(x_k,v_i) \tag{11}$$

其中，c 为类别数，n 为数据向量数，x_k 为第 k 个样本向量，v_i 为第 i 类中心向量，$\sum_{i=1}^{c}u_{ik}=1,u_{ik}\in(0,1),\forall k,d(x_k,v_i)=\|x_k-v_i\|^2$。为了便于发现离群点，为每一个样本点分配一个动态权值，所得新的目标优化迭代函数如下：

$$J_m(U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\frac{1}{w_k^q}d^2(x_k,v_i) \tag{12}$$

其中因子 w_k 表示第 k 个样本的权值，同时 $\sum_{k=1}^{n}w_k=w$，w 为用户定义的一个实值常数。将输入样本数据经核的 PP 主成分数据变换投影到特征空间的 $P(x_k)$，带入(12)式得新的目标

优化迭代函数(13)式：

$$J_{mG}(X,U,v)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\frac{1}{w_k^q}d^2(P(x_k),\overline{v_i}) \tag{13}$$

其中$\overline{v_i}$表示特征空间中第 i 类的中心。

离群聚类算法通过常实值参数 q 来控制离群权值的影响。离群聚类算法的目的是用 $1/w_k^q$ 表示一个动态权值，它表示数据样本离群程度。给普通数据样本分配一个小权值 $w_k(1/w_k^q$ 大)；而一般地，离群点通常远离任何一个类中心，从而分配一个大权值 $w_k(1/w_k^q$ 小)参数 q 在离群聚类过程中起到动态调控作用：当 q 足够大时，每个数据样本的权值将趋近相等 w/k，也就是说，权值对于所有的样本都有相同的影响；当 $q\to 0$ 时，权值影响将达到最大。

当 w 为常数时，假定隶属度固定，则使用 Lagrange 乘子寻优算法导出目标优化迭代函数的最小值，可获得 w_k 的迭代式：

$$J_{mG}(X,U,\overline{v})=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\frac{1}{w_k^q}d^2(P(x_k),\overline{v_i})+\lambda(\sum_{k=1}^{n}w_k-w) \tag{14}$$

对(14)式中的 w_k 求偏微分获得(15)式

$$\frac{\partial J_{mG}(X,U,\overline{v})}{\partial w_k}=-q*\frac{1}{w_k^{q+1}}\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i})+\lambda \tag{15}$$

令$\frac{\partial J_{mG}(X,U,\overline{v})}{\partial w_k}=0$，则可得(16)式

$$\lambda=q*\frac{1}{w_k^{q+1}}\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}) \tag{16}$$

由(16)式可得(17)式

$$w_k=\left(\frac{q\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}}{\lambda}\right)^{\frac{1}{q+1}} \tag{17}$$

根据$\sum_{k=1}^{n}w_k=w$，由上式可得(18)式

$$w=\sum_{k=1}^{n}\left(\frac{q\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i})}{\lambda}\right)^{\frac{1}{q+1}} \tag{18}$$

且因此得(19)式和(20)式

$$\frac{1}{\lambda^{q+1}}=\sum_{k=1}^{n}(q\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}))^{\frac{1}{q+1}}*\frac{1}{w} \tag{19}$$

$$\lambda=\left(\sum_{k=1}^{n}(q\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}))^{\frac{1}{q+1}}*\frac{1}{w}\right)^{q+1} \tag{20}$$

由(16)式和(20)式联立可最终得到 w_k 的迭代公式(21)

$$w_k=\frac{(q\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}))^{\frac{1}{q+1}}}{\sum_{k=1}^{n}(q\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}))^{\frac{1}{q+1}}}*w$$

$$=\frac{(\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}))^{\frac{1}{q+1}}}{\sum_{k=1}^{n}(\sum_{i=1}^{c}(u_{ik})^m d^2(P(x_k),\overline{v_i}))^{\frac{1}{q+1}}}*w \tag{21}$$

同样在满足约束条件$\sum_{i=1}^{c}u_{ik}=1,\forall k=1,\cdots,n$ 情况下，使用 Lagrange 乘子寻优算法可获得隶属度和类中心的迭代式：

$$u_{ik}=\frac{1}{\left(\sum_{j=1}^{c}\frac{d^2(P(x_k),\overline{v_i})}{d^2(P(x_k),\overline{v_j})}\right)^{\frac{1}{m-1}}} \tag{22}$$

$$v_i=\frac{\sum_{k=1}^{n}(\overline{u_{ik}})^m P(x_k)}{\sum_{k=1}^{n}(\overline{u_{ik}})^m} \tag{23}$$

其中，$\overline{u_{ik}^m}=\frac{u_{ik}^m}{w_k^q}$

基于上面的分析，我们建立离群聚类算法，从非线性降维中，得到相应样本数据的权值，进而确定样本数据集中的离群数据点，其工作过程说明如下：

步骤 1　对算法参数 C、q、m、ε 初始化，其中 C 为聚类数目，q 为权值指数，m 为模糊因子，ε 是一个很小的正数，算法的迭代次数记为 s；

步骤 2　由基于核的 PP 主成分数据变换得到投影函数 $P(x)$；

步骤 3　将权值及投影函数 $P(x)$ 带入(21)式中，对数据集中样本数据加权后建立新的目标优化迭代函数 J；

步骤 4　初始化样本数据集中样本隶属度参数，同时初始化样本向量到类中心的距离；

步骤 5　根据初始化中心距离重新计算每一样本向量的隶属度$(u_{ik})^s$；

步骤 6　对样本数据进行检测，根据上面的公式(21)生成每个样本向量权值；

步骤 7　重复步骤 5 和步骤 6，直到前后两次迭代所得结果趋向稳定，从而完成离群数据点的发现。这里取当第 s 次迭代与第 $s+1$ 次迭代产生的$|J^{s+1}-J^s|<\varepsilon$ 时停止迭代。

4　离群聚类算法收敛性分析

定理　离群聚类算法的目标优化迭代函数 J 中，$u_{ik}(I=1,2\cdots,c,k=1,2,\cdots,n)$和 $w_k(k=1,2,\cdots,n)$是 J 局部最小值的必要条件是 $u_{ik}=\frac{1}{\left(\sum_{j=1}^{c}\frac{d^2(P(x_k),\overline{v_i})}{d^2(P(x_k),\overline{v_j})}\right)^{\frac{1}{m-1}}}$ 满足$\sum_{i=1}^{c}u_{ik}=1$ 且由式(21)所得 w_k 满足约束条件$\sum_{k=1}^{n}w_k=w$。

证明：首先假设给定 w_k，则问题转变为求目标优化迭代函数 J 对 u_{ik} 的最小值并满足约束条件$\sum_{i=1}^{c}u_{ik}=1$，由 Lagrange 函数可获得下式：

$$L(W,\lambda)=J-\sum_{k=1}^{n}\lambda_k(\sum_{i=1}^{c}u_{ik}-1) \tag{24}$$

对(24)式的 u_{ik} 和 λ_k 求偏微分可获得(25)式和(26)式：

$$\frac{\partial L(W,\lambda)}{\partial u_{ik}}=m(u_{ik})^{m-1}\frac{1}{w_k^q}d^2(P(x_k),\overline{v_i})-\lambda_k=0 \tag{25}$$

$$\frac{\partial L(W,\lambda)}{\partial \lambda_k}=\sum_{i=1}^{c}u_{ik}-1=0 \tag{26}$$

由(25)可解得(27)式：

$$u_{ik}=\left[\frac{\lambda_k w_k^q}{md^2(P(x_k),\overline{v_i})}\right]^{\frac{1}{m-1}} \tag{27}$$

将(27)式代入(26)式可得(28)式：

$$\left[\frac{\lambda_k w_k^q}{m}\right]^{\frac{1}{m-1}}=\frac{1}{\left(\sum_{i=1}^{c}\frac{1}{d^2(P(x_k),\overline{v_i}}\right)^{\frac{1}{m-1}}} \tag{28}$$

将(28)式代入(27)式可得隶属度迭代函数式(22)。要证明权值 w_k 由(21)式获得，并满足约束条件$\sum_{k=1}^{n}w_k=w$，其证明过程可依据 w_k 的迭代公式推导得到，此处从略。

对于离群聚类算法的目标优化迭代函数 J 而言，显然迭代序列是有下界的，因为目标函数值及其隶属度均不会小于零。从任意一步(U^s,W^s)到下一次迭代(U^{s+1},W^{s+1})，算法是经过两次交互的寻优过程，即从(U^s,W^s)到(U^{s+1},W^s)，再从(U^{s+1},W^s)到(U^{s+1},W^{s+1})。由上述定理可知，在前一步，因为 $J(U^s,W^s)$是目标函数在(U^s,W^s)处的极小值点，所以有 J

$(U^s,W^s)\geqslant J(U^{s+1},W^s)$。同理，因为 $J(U^{s+1},W^{s+1})$ 是目标函数在 (U^{s+1},W^s) 的极小值点，所以有 $J(U^{s+1},W^s)\geqslant J(U^{s+1},W^{s+1})$。综合上述两个步骤可得 $J(U^s,W^s)\geqslant J(U^{s+1},W^s)\geqslant J(U^{s+1},W^{s+1})$。因此，可以证明目标优化迭代函数 J 存在局部最优解，使得 $J(U^{s+1},W^{s+1})\leqslant J(U^s,W^s)$，即目标优化迭代函数 J 是 s 的递减函数，所以本文提出的离群聚类算法最终收敛。

5 算法的实现及其实验结果的评估

5.1 实验数据集与预处理

本节通过实验验证算法的有效性、效率并对离群点发现质量进行分析。测试数据集信息采用基于网络的安全审计数据，数据来源于美国国防部高级研究计划署(PARDA)在1998年由麻省理工学院 Lincoln 实验室提供的用于入侵检测系统评估的数据。该数据集源于美国空军局域网的仿真环境，每个实例包含 42 个属性，这些数据中包含特定的攻击模式，均已标识为正常或特定的攻击行为。数据集中入侵类型按攻击手段类型可划分为：拒绝服务攻击(DoS)；远程权限获取(R2L)；各种权限提升(U2R)；各种端口扫描和漏洞扫描(Probe)。由于原始数据集过于庞大，且分布不均匀，本实验使用其中的部分数据，并对非数值属性值做了预处理，经过预处理构成实验数据集，其为高维空间的数据点，数据集中所包含的各种攻击数量见图 1。首先由基于核的 PP 主成分数据变换获得包括离群点的二维平面样本数据，参见图 2(包括离群点的线性可分样本数据)和图 3(包括离群点的线性不可分样本数据)。

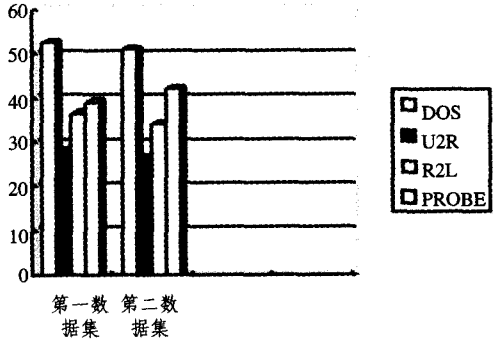

图 1 数据集中各种攻击类型的数量

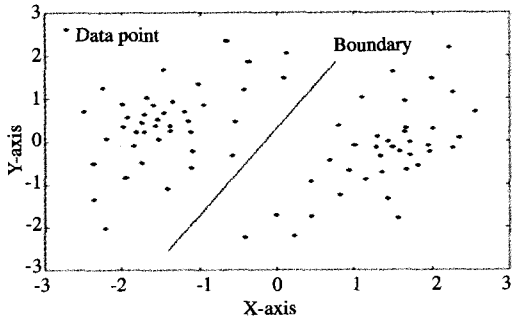

图 2 包括离群点的线性可分的样本数据集

5.2 实验结果及评估

首先给出离群聚类算法的参数选择，然后提供该算法对线性可分和线性不可分样本数据集应用的比较情况。

参数选择：离群聚类算法共有三个参数需要确定，分别为模糊因子 m、权值指数 q 及用户定义常量 w。模糊因子 m 的最佳值区间在[1.5,2.5]范围内[15]，本文选取常用值 $m=2$，权值指数 $q=1$ 及用户定义常量 $w=200$。

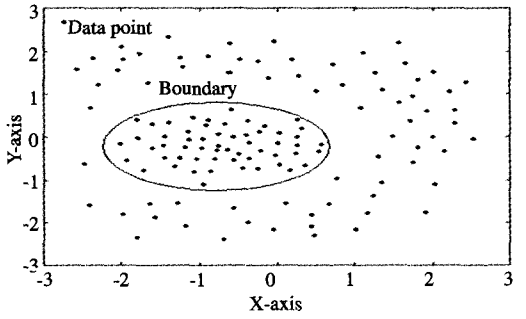

图 3 包括离群点的线性不可分样本数据集

第一个实验重复进行了 1000 次。图 4 为每次聚类收敛到发现离群点的迭代次数。图中的横坐标 n 表示第 n 次实验，纵坐标 I 表示收敛到发现最远离群点时的迭代次数。由图 4 可见，算法通常在 4 次迭代之内发现离群点，在 10 次以上发现离群点的概率要小于 1‰。由图 5 可见，算法在聚类过程中发现离群点的个数随迭代次数的变化。

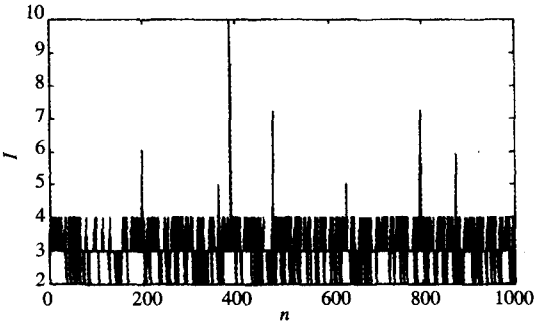

图 4 对图 2 样本数据集聚类时发现离群点的迭代次数

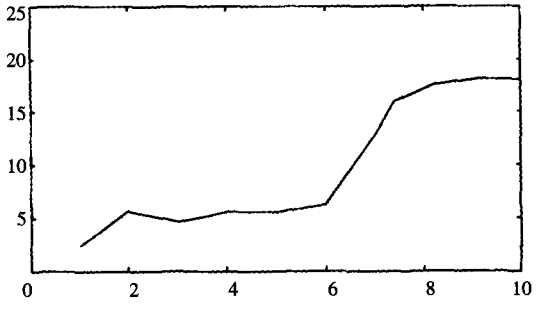

图 5 本文提出的算法在聚类过程中发现离群点的个数随迭代次数的变化(横坐标为迭代次数，纵坐标为离群点个数)

对于图 3 中线性不可分样本数据集，从图 6 中可以看出，当算法结束时，根据最终得到的权值，可以作出等权值线，在一定的等权值线之外的数据点就被认为是该样本集在该权值下的离群点，权值的选择由专家给出。对于线性可分的数据样本集，本文提出的离群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[2～6]相似的结果，但从结果看出，本文所提出的新算法权值变化幅度较小，也就是说，对待离群点更慎重，而这一点更符合人

(下转第 138 页)

结论 针对标准人工免疫算法中存在的早期收敛和后期收敛速度慢的问题，提出了一种基于小生境技术和聚类分析的人工免疫算法。利用嵌入进化标记的小生境技术，不仅加快了算法的收敛速度，而且保持了种群的多样性，有效地防止早熟。当种群进化到一定程度后，进行了聚类分析，使得人工免疫算法的异化算子在类与类之间可以进行全局搜索，即运用粗搜索找到适应度较高的区域，再用趋同算子在类内抗体中进行局部搜索以寻求高精度的解。算法通过粗和细的两层搜索，保证了全局寻优的速度和精度。实验表明，该算法是一种寻优能力强、效率和可靠性更高的优化算法，其综合性能比标准的人工免疫算法有了一定的提高。

参 考 文 献

1 王磊，潘进，焦李成. 免疫算法[J]. 电子学报，2000，28(7)：74～78
2 Vargas P A, Castro L N, Von Zuben F J. Artificial immune systems as complex adaptive systems. In: Proceedings of International Conference on Artificial immune systems, 2002, 1: 115～123
3 Miller B L, Shaw M J. Genetic Algorithms with Dynamic Niche Sharing for multimodal Function Optimization. In: IEEE International Conference on Evolutionary Computation, Piscataway, NJ: IEEE Press, 1996, 786～791
4 Goldbergde, Richardson J. Genetic Algorithm with Sharing for Multimodal Function Optimization[A]. In: Proceedings of 2nd International Conference on Genetic Algorithm, Lawrence Erlbaum Associates[C], 1987. 41～49
5 Jain A K, Dubes R C. Algorithms for clustering[M]. Englewood Cliffs: N J Prentice Hall, 1988

(上接第 134 页)

们对待离群点的正常思维方式。对于线性不可分的数据样本集，本文提出的离群聚类算法是在引入非线性函数，映射至特征空间进行聚类以及根据等权值线标识离群点的结果，所以能够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，因此能够更容易地发现样本数据集中的离群点。

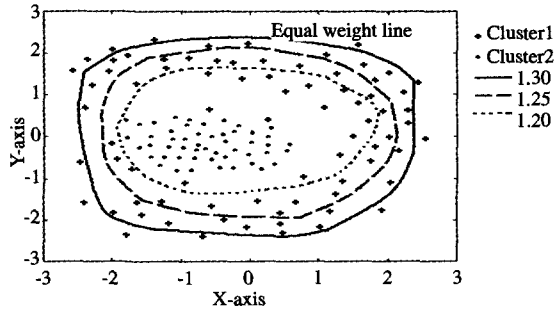

图 6 本文提出的算法对线性不可分样本数据集聚类的结果

对实验结果分析发现：

①实验结果与样本数据点的个数和权值有很大关系，选取不同数目的样本数据点和分配所得的不同权值对边缘点的影响都会比较大；由于离群点通常远离任何一个类中心，从而分配一个大权值 w_k($1/w_k^q$ 小)。权值指数 q 等于 1.5 时，本算法就是基于重心的算法，而权值指数 q 趋近 0 时，样本数据点取所有点时，本算法就是基于所有样本数据点的算法。经过测试权值指数 q 选取 1 时效果较好。②因为测试数据选取有限，所以没能进行更广泛的测试，通过观察分类结果，发现其中有很多是分散样本数据点归为一类，而一类最多包含的样本数据又过多，可见每一类所包含文本数不太平衡，说明在许多方面还有待改进。

结束语 本文在基于核的 PP 主成分的基础上提出一种离群聚类算法。该算法首先用数据变换对输入数据进行预处理，数据预处理的一个重要结果就是得到降维的投影向量。通过引入非线性变换函数，把预处理所得数据映射到特征空间，通过为特征空间中的向量分配一个动态的权值，最终在优化经典的 FCM 目标优化迭代函数基础上，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数据集，在运用传统算法失败的情况下，本文提出的离群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。

该算法优点包括：

a 能够有效地解决高维空间中数据的稀疏问题；

b 能够找到合适的衡量办法给出高维子空间中离群点的物理意义；

c 适用于线性不可分的数据集；

d 对高维空间中的数据仍然是计算高效的。

最后，我们也证明了离群聚类算法的收敛性。仿真实验表明了该算法的有效性。

参 考 文 献

1 夏火松主编．数据仓库与数据挖掘技术．北京：科学出版社，2004
2 Barnett V, Lewis T. Outliers in Statistical Data[M]. New York: John Wiley and Sons, Inc., 1994
3 Franco P, Ian S M. Computational Geometry: an Introduction [M]. New York: Springer-Verlag, 1988
4 Knorr Edwin M, Ng R T. Algorithms for Mining Distance-Based Outliers in Large Datasets[R]. In: Proceedings of the 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann, 1998. 392～403
5 Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying Density-based Local Outliers[R]. In: Chen W, Naughton J F, Bernstein P A, eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas: ACM Press, 2000. 93～104
6 Spiros P, Hiroyuki K, Gibbons P B. LOCI: Fast Outlier Detection Using the Local Correlation Integral[R]. In: Proceedings of the 19th International Conference on Data Engineering, 2003. 315～326
7 Beyer K, Goldstein J, Ramakri S R, et al. When is nearest neighber meaningful? In: Been, C, Buneman P, eds. Proceedings of the 7th Intemational Conference 0n Data Theory Lecture Notes In Computer Science 1 540. Jernsalem: Spnnger, 1999. 217～235
8 Li Y. Reforming the theory of invariant moments for pattern recogoition[J]. Pattern Recognition, 1992, 25(7): 723～730
9 魏藜，钱卫宁，周傲英. SLOT：基于估计的高效子空间局部离群点发现[J]. 计算机科学，2002，29(8)：122～125
10 Hinneburg A, Aggarwal C C, Keim D A. What is the nearest neighbors in high dimensional space? [n]. In: Abbadl A E, Brodie M L, Chakravarthy S, et al, eds. Proceedings of the 26th International Conference on Very Large Data Bases Cairo: Morgan Kaufmann, 2000. 506～515
11 项静恬，史久恩. 非线性系统中数据处理的统计方法．北京：科学出版社，1997. 171～194
12 Friedman J H, Tukey J W. A projection pursuit algorithm for exploratory data analysis[J]. IEEE Tram Computer, 1974, 23(9): 881～889
13 Suykens J A K, Gestel T V, Vandewalle J, et al. A Support Vector Machine formulation to PCA Analysis and its Kernel version[R]. [ESAT-SCD-SISTA Technical Report 2002-68]. Belgium: Katholieke Universiteit Leuven, 2002
14 Suykens J A K, Gestel T V, Vandewalle J, et al. A Support Vector Machine formulation to PCA Analysis and its Kernel version[R]. [ESAT-SCD-SISTA Technical Report 2002-68]. Belgium: Katholieke Universiteity Leuven, 2002
15 Pal N R, Bezdek J C. On cluster validity for the fuzzy c-mesns model[J]. IEEE Trans Fuzzy System, 1995, 3(3): 370～379