

切换回归模型的抗噪音聚类算法

杨小兵¹, 何灵敏¹, 孔繁胜²

(1. 中国计量学院 计算机系, 浙江 杭州 310018; 2. 浙江大学 计算机学院, 浙江 杭州 310012)

摘要:对切换回归模型的聚类方法一般都没有考虑到噪音的影响,因此在含有噪音数据的情况下,用这些方法聚类的结果就会出现一定的偏差.为了减弱聚类过程中噪音数据的影响,提出了一种新的具有抵抗噪音能力的聚类算法,称为抗噪音聚类算法.该算法通过将已知数据集划分为非噪音数据集和噪音数据集2个子集,然后对非噪音数据集进行聚类分析,估计出模型的各个参数.通过对噪音数据集和非噪音数据集进行不断地调整,同时不断地修正得到的参数估计值,从而得到对聚类结果的优化.实验表明,抗噪音聚类算法能够有效地克服噪音数据对聚类结果的影响,并估计出优质的参数.

关键词:切换回归模型;聚类;抗噪音聚类算法

中图分类号:TP301.6 **文献标识码:**A **文章编号:**1673-4785(2009)06-0497-05

A noise-resistant clustering algorithm for switching regression models

YANG Xiao-bing¹, HE Ling-min¹, KONG Fan-sheng²

(1. Department of Computer Science, China Jiliang University, Hangzhou 310018, China; 2. College of Computer Science, Zhejiang University, Hangzhou 310012, China)

Abstract: Clustering methods for switching regression models usually neglect the effects of noise. As a result, errors usually exist if clustering is carried out in a noisy environment. In order to overcome the effects of noise, a new clustering algorithm, a noise-resistant clustering algorithm, was proposed. The algorithm partitions the dataset into two sub-datasets, a noiseless dataset and a noisy dataset, and then performs clustering analysis on the noiseless dataset to estimate parameters. By continuous simultaneous adjustment of the noisy and noiseless datasets and by continuously revising estimated parameters, the results of clustering were improved. Simulation experiments demonstrated that the algorithm efficiently clusters noisy datasets and can provide good estimates of parameters.

Keywords: switching regression models; clustering; noise-resistant clustering algorithm

聚类分析是当前研究的一个热点,已经广泛地应用在模式识别、数据分析、图像处理等许多领域中^[1].混合分布模型是一种由分别满足多种不同分布规律的数据混合在一起的模型,是统计学中最重要的模型之一,这种模型也已经应用到聚类问题的研究中.由于参数回归往往能够很好地描述数据集的特征,因此参数回归模型成为对数据集进行聚类分析的一种重要形式.通常要分析的数据集并不是满足一个简单的回归模型,而是有多个简单回归模型混合而成的混合回归模型,即切换回归模型(switching regression models),切换回归模型就是一

种特殊的混合分布模型.

1 切换回归模型

假设 $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ 是一个数据集,其中 $x_k \in \mathbf{R}^s, y_k \in \mathbf{R}^t$. 在最简单的数据回归问题中,假设 x 和 y 满足一个简单的函数关系,即

$$y = f(x; \beta) + \varepsilon. \quad (1)$$

式中: $\beta \in \Omega \subset \mathbf{R}^k$ 为待确定的参数向量, ε 为随机向量,其均值向量为 $\mu = 0 \in \mathbf{R}^t$,协方差矩阵为 Σ . 然而,在很多情况下, x 和 y 并不是简单地满足函数关系式(1),而是满足一种更复杂的函数关系:

$$y = f_i(x; \beta_i) + \varepsilon_i, 1 \leq i \leq c. \quad (2)$$

其中每个参数向量 $\beta \in \Omega \subset \mathbf{R}^{k_i}$, 每个随机向量 ε_i 对应的均值向量和协方差矩阵分别为 $\mu_i = 0 \in \mathbf{R}^t$ 和

Σ_i . 基于这种函数关系的模型称为切换回归模型. 切换回归模型在很多领域尤其是经济领域^[2,4]已经得到广泛的应用.

一个典型的例子可以用来解释这种模型^[5]. 人们在对鱼类的研究中发现, 一种大比目鱼有种有趣现象, 在一定的年龄范围内, 雄性的大比目鱼的平均长度和它的年龄近似地满足一个线性函数关系; 同样, 雌性的大比目鱼的平均长度和它的年龄近似地满足另一个线性函数关系, 这2个线性函数关系并不相同. 大比目鱼的长度和年龄是可以测量或得到的, 而它们的性别是未知的. 对这2个线性函数的参数的确定问题就是一个简单的切换回归问题, 这里 $c=2, s=1, t=1, y$ 为长度, x 为年龄, 这个切换回归模型就可以表示为

$$\begin{cases} y = f_1(x, \beta_1) + \varepsilon_1 = \beta_{11}x + \beta_{12} + \varepsilon_1, \\ y = f_2(x, \beta_2) + \varepsilon_2 = \beta_{21}x + \beta_{22} + \varepsilon_2. \end{cases} \quad (3)$$

假设对不同的数据 (x_k, y_k) 和 (x_j, y_j) 来说, ε_1 和 ε_2 的值总是相互独立的, 并且分别满足均值为0, 方差分别为 σ_1 和 σ_2 的正态分布. 目前已经有一些方法对切换回归模型进行参数估计, 如硬 c -平均算法 (hard c -means, HCM)^[6] 和模糊 c -回归模型算法 (fuzzy c -regression models, FCRM)^[7] 等.

2 已有算法及其存在的缺陷

含有 c 个回归函数分支的切换回归模型又可称为 c -回归模型. 要对 c -回归模型的参数进行估计, 一个简单的思路就是, 首先将数据集 S 划分为 c 个子集, 然后对每个子集分别进行单模型回归, 从而得到每个子函数的参数. 这种方法称为硬 c -划分法. 硬 c -划分法根据划分方法的不同而有多种算法, 比如硬 c -平均算法. 这类算法虽然实现起来比较容易, 但由于它们没有考虑到各个回归函数分支的重叠的情况, 即没有考虑到数据的模糊性, 因此它们的应用存在一定的局限性.

Hathaway 和 Bezdek 在 1993 年提出了一种模糊 c -回归模型算法^[7], 这种算法在对数据集 S 进行模糊 c -划分的同时估计 c -回归模型的参数. 对于回归模型(1)来说, 隶属度 (membership degree) U_{ik} (U_{ik} 表示隶属度矩阵 U 的第 i 行第 k 列的值, $0 \leq U_{ik} \leq 1$ 且 $\sum_{i=1}^c U_{ik} = 1$) 用来表示模型 $f_i(x_k; \beta_i)$ 在多大程度上与 y_k 匹配. 误差判定标准可以是各种测量 $f_i(x_k; \beta_i)$ 与 y_k 之间的误差的方法, 最常用的是

$$E_{ik}(\beta_i) = \|f_i(x_k; \beta_i) - y_k\|^2. \quad (4)$$

FCRM 的目标函数族可以定义为

$$E_m(U, \{\beta_i\}) = \sum_{k=1}^n \sum_{i=1}^c U_{ik}^m E_{ik}(\beta_i). \quad (5)$$

式中: $m > 1$ 为一个决定结果簇的模糊性的加权指数.

FCRM 算法通过最小化目标函数族(5)来估计每个参数 β_i 和隶属度 U_{ik} 的值. FCRM 算法在无噪音的情况下可以广泛应用于切换回归模型的参数估计和聚类, 并能够得到较好的计算结果.

近年来, 通过对 FCRM 算法的研究, 一些改进的方法被提了出来, 如 PPSR 算法^[8-9]、离群模糊核聚类算法^[10]、GFC 算法^[11-12] 等. 这些算法增强了 FCRM 算法的功能, 扩大了 FCRM 算法的应用范围, 然而, 它们和 FCRM 算法一样忽略了噪音数据的干扰. 当数据集中含有噪音数据尤其是大量噪音数据时, 不管是 HCM 算法还是 FCRM 算法及其一些改进方法, 都无法排除噪音的干扰, 计算结果也因此会受到影响. 由于这些算法没有考虑噪音数据的影响, 把它们统称为无噪音聚类算法 (no noise clustering algorithm). 为了能够更有效地排除数据集中的噪音数据的干扰, 提出了一种以无噪音聚类算法为基础, 但又与选用何种无噪音聚类算法无关, 并且具有抵抗噪音能力的聚类算法——抗噪音聚类算法 (noise resistance clustering algorithm, NRC Algorithm). 该算法能够对数据集中存在的大量噪音数据进行清除, 有效地克服了噪音数据对计算结果的影响, 很好地解决了数据集中的噪音问题.

3 抗噪音聚类算法

在切换回归模型(2)中, 为了识别出数据集中的噪音数据, 用 $d_{ki}(x_k, y_k)$ 表示数据点 (x_k, y_k) 到 $y = f_i(x; \beta_i)$ 的距离. 这里的距离是一个抽象的概念, 它是用来判断数据点 (x_k, y_k) 与 $y = f_i(x; \beta_i)$ 的相似程度的标准, 例如在二维空间中, 它可以用点 (x_k, y_k) 到曲线 $y = f_i(x; \beta_i)$ 的最短距离表示, 也可以用点 (x_k, y_k) 到曲线 $y = f_i(x; \beta_i)$ 中纵坐标与 (x, y_k) 相同的点的距离, 即 $|y_k - f_i(x_k, \beta_i)|$ 表示, 可见它是一个相对的标准. 设定一个距离阈值 w , 当 $d_{ki}(x_k, y_k) \leq w$ 时, 说明与 $y = f_i(x; \beta_i)$ 之间的距离足够近, 可以判断出它不是噪音数据, 或者即使它是由于噪音产生的数据也不会对回归结果产生负面影响, 因此可以认为 (x_k, y_k) 不是噪音数据; 而当 (x_k, y_k) 对所有的 $y = f_i(x; \beta_i)$ ($1 \leq i \leq c$) 的距离都大于 w 时, 就可以认为 (x_k, y_k) 是一个噪音数据.

根据以上思想, 提出了切换回归模型的抗噪音聚类 (noise resistance clustering algorithm, NRC), 算法的流程图如图 1 所示.

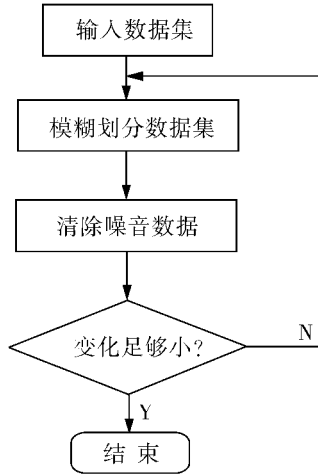


图1 NRC算法的流程图

Fig. 1 Flow chart of NRC algorithm

假设 S_{noise} 是数据集 S 中的噪音数据的集合,相应地 S_{data} 是非噪音数据的集合,那么 NRC 算法的目的就是将 S 分为 2 个子集 S_{data} 和 S_{noise} ,然后对 S_{data} 进行无噪音聚类,估计出各个子模型及对应的参数。

在 NRC 算法中,首先 S_{data} 的初始值为 S , S_{noise} 初始值为 \emptyset . 然后利用无噪音聚类算法将 S_{data} 划分为 c 个子集并回归为 c 个单模型,同时估计出各个子模型的参数 β_i ($1 \leq i \leq c$). 在 S_{data} 含有噪音的情况下,通过无噪音聚类算法得到的划分和参数是不准确的,因此需要对它们进行修正. 修正的方法就是根据设定的距离阈值 w 对 S_{data} 中的每个数据点 (x_k, y_k) 进行判断,如果 (x_k, y_k) 被认为是噪音的话就将其从 S_{data} 去除,然后加入到 S_{noise} 中;同时随着算法的执行,参数发生变化时也要对 S_{noise} 中原来认为是噪音的数据重新进行判断,如果有数据点到某个子模型的距离小于 w 了,就把它在 S_{noise} 中去除加入到 S_{data} 中. 以上过程循环执行,直到 S_{data} 和 S_{noise} 不再发生变化. NRC 算法的一般步骤可以描述为

1) 已知 $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$. 确定模型 (2) 并根据实际应用的需要设定距离范围 w , 并选择终止循环参数 $\varepsilon > 0$. 初始化非噪音数据集 $S_{\text{data}}^{(0)} = S$ 和噪音数据集 $S_{\text{noise}}^{(0)} = \emptyset$, 同时令 $r = 0$.

2) 利用无噪音聚类算法对 $S_{\text{data}}^{(r)}$ 进行划分回归和估计参数,得到各子模型的参数值 $\beta_i^{(r)}$ 及模型 $y^{(r)} = f_i(x; \beta_i^{(r)}) + \varepsilon_i$, $1 \leq i \leq c$.

3) 更新 $S_{\text{data}}^{(r)} \rightarrow S_{\text{data}}^{(r+1)}$: 扫描 $S_{\text{data}}^{(r)}$ 中的数据,对于其中的每一个数据点 (x_k, y_k) , 根据它被划分到的子集所对应的子模型 $y^{(r)} = f_i(x; \beta_i^{(r)}) + \varepsilon_i$, 计算 $d_{ki}(x_k, y_k)$, 如果 $d_{ki}(x_k, y_k) > w$, 则把 (x_k, y_k) 从 $S_{\text{data}}^{(r)}$ 清除, 加入到 $S_{\text{noise}}^{(r)}$ 中.

4) 更新 $S_{\text{noise}}^{(r)} \rightarrow S_{\text{noise}}^{(r+1)}$: 扫描 $S_{\text{noise}}^{(r)}$ 中的数据,对于

其中的每一个数据点 (x_k, y_k) , 如果它不是刚从 $S_{\text{data}}^{(r)}$ 中转移过来的, 则对 $0 \leq i \leq c$ 分别计算 $d_{ki}(x_k, y_k)$. 如果存在某个 $d_{kj}(x_k, y_k) \leq w$ ($1 \leq j \leq c$), 那么就把 (x_k, y_k) 从 $S_{\text{noise}}^{(r)}$ 清除, 加入到 $S_{\text{data}}^{(r)}$ 中.

5) 如果 $S_{\text{data}}^{(r)}$ 和 $S_{\text{noise}}^{(r)}$ 不再发生变化或者 $\|\beta^{(r)} - \beta^{(r+1)}\| < \varepsilon$, 则程序终止, 返回 $S_{\text{data}}^{(r)}$ 和 $S_{\text{noise}}^{(r)}$, 以及对应的模型 $y^{(r)} = f_i(x; \beta_i^{(r)}) + \varepsilon_i$, $1 \leq i \leq c$ 和参数 $\beta_i^{(r)}$; 否则令 $r = r + 1$, 转到 2).

由于给定的数据集 S 中的数据有限性, 算法必然会在有限步内终止. 在实际执行过程中, 第 2) 步所采用的无噪音聚类算法可以是简单易行的硬划分方法, 如 HCM 算法, 也可以是考虑更为合理的模糊划分算法, 如 FCRM 算法. 如果是采用了模糊划分算法, 那么在判断 S_{data} 中的点 (x_k, y_k) 是否为噪音数据时, 可以以 (x_k, y_k) 所对应的隶属度最大的子模型对应的方程作为判断的参照方程. 即如果 U_{jk} 是 U_{ik} ($1 \leq i \leq c$) 中的最大值, 那么就以 $d_{ki}(x_k, y_k)$ 与 w 的比较来作为判断 (x_k, y_k) 是否是噪音数据的标准.

由于 NRC 算法的每次迭代都是在去除噪音数据之后对模型的参数进行估计的, 因此通过其得到的参数将会明显优于其他未考虑噪音数据的聚类算法, 仿真实验将进一步证明这一点.

4 仿真实验

为了证明 NRC 算法对噪音数据的抵抗能力, 给出 2 个仿真实验, 实验中采用的模型都为 $c = 2$ 的切换回归模型, 其中第 1 个实验用来估计线性回归模型的参数值, 第 2 个实验用来估计二次回归模型的参数值.

实例一:

在本实验中用到的实验数据为 600 个包含有大量噪音数据的数据点, 其中的非噪音数据都是按模型 (3) 生成的. 模型的参数分别设为 $\beta_{11} = 0, \beta_{12} = 0, \sigma_1 = 0.25, \beta_{21} = 1.0, \beta_{22} = 0$ 和 $\sigma_2 = 0.25$. 每个数据点 (x_k, y_k) 的生成过程如下:

1) 按均匀分布随机生成一个数 $z_k \in (0, 1)$, 如果 $z_k < 0.35$, 则令 $i = 1$, 如果 $0.35 \leq z_k < 0.7$, 则令 $i = 2$, 否则令 $i = 0$;

2) 在 $(-3, 3)$ 上按均匀分布随机选定 x_k , 并按均值为 0, 方差为 σ_i 的正态分布随机生成误差 ε_i ;

3) 按如下方法计算 y_k : 如果 $i = 0$, 则在 $(-3, 3)$ 上按均匀分布随机生成 y_k , 这时 (x_k, y_k) 就是噪音数据, 否则根据模型 (3), 通过 $x_k, \varepsilon_i, \beta_{i1}$ 和 β_{i2} 计算 y_k , 这时的 (x_k, y_k) 为非噪音数据.

从数据的生成过程可以看出这个切换回归模型

就是 $f_1(x) = 0$ 和 $f_2(x) = x$. 实验数据的示意图如图 2 所示.

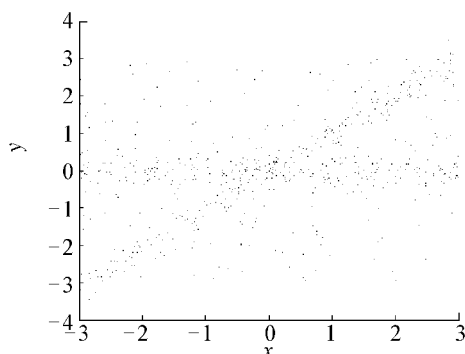


图 2 实例 1 生成的数据集的示意图

Fig. 2 Scatter plot of simulation data for example 1

随后对上面生成的数据集分别执行无噪音聚类算法和 NRC 算法. 在实验中采用的无噪音聚类算法是 FCRM, 由于 FCRM 算法是一种模糊算法, 为了便于比较, 对每个数据点按隶属度最高原则将其划分到具体的子集中. 在 NRC 算法中, 选取的距离阈值为 $w = 1.0$.

聚类结果如图 3 和图 4 所示.

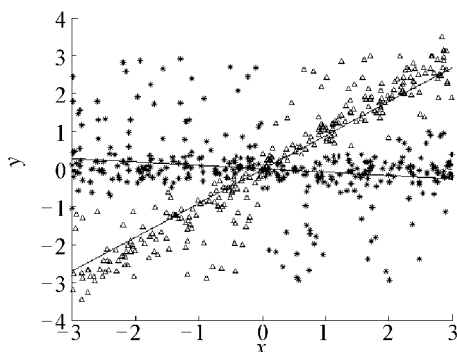


图 3 实例 1 无噪音聚类算法聚类结果

Fig. 3 Clustering results of no noise clustering algorithm for example 1

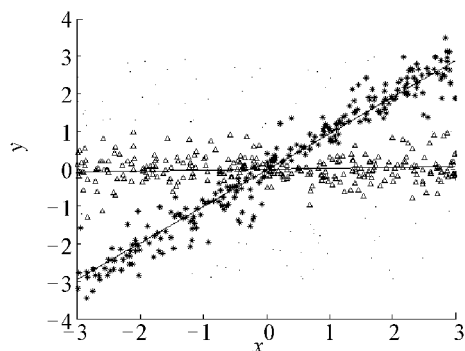


图 4 实例 1 NRC 算法聚类结果

Fig. 4 Clustering results of noise resistance clustering algorithm for example 1

图 3 与图 4 中的“*”和“Δ”分别表示划分结果的 2 个不同的子集, 直线分别表示 2 个算法所得到的回归模型. 其中图 4 中的“.”表示 NRC 算法识别出来的噪音数据.

FCRM 算法得到的模型参数分别为 $\beta_{11} = -0.0896$, $\beta_{12} = 0.0091$, $\beta_{21} = -0.8968$, $\beta_{22} = -0.0116$, NRC 算法得到的模型参数分别为 $\beta_{11} = -0.0243$, $\beta_{12} = -0.0094$, $\beta_{21} = 0.9739$, $\beta_{22} = -0.0358$. 通过实验结果可以看出, NRC 算法得到的聚类效果明显要比 FCRM 算法得到的聚类效果好.

实例二:

在这个实验中, 将利用 NRC 算法来估计二次回归模型的参数, 并与用 FCRM 算法的估计结果相比较. 实验选取的模型为

$$\begin{cases} y = 3 - \frac{2}{3}x^2, \\ y = 3 + \frac{2}{3}x^2. \end{cases} \quad (6)$$

用类似实例一的方法生成 600 个数据点, 如图 5 所示.

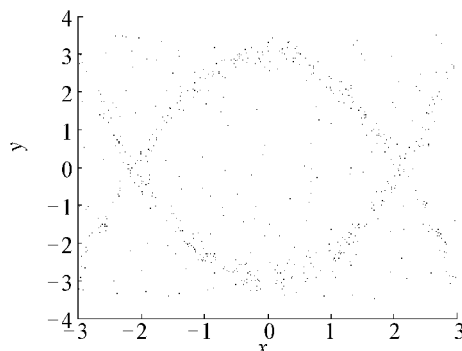


图 5 实例 2 生成的数据集的示意图

Fig. 5 Scatter plot of simulation data for example 2

对生成的数据分别应用无噪音聚类算法和 NRC 算法后得到的参数估计值如表 1 所示, 聚类结果如图 6 和图 7 所示.

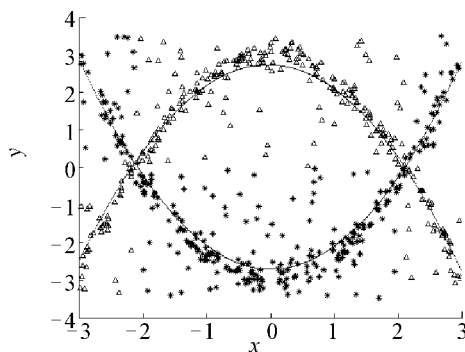


图 6 实例 2 无噪音聚类算法聚类结果

Fig. 6 Clustering results of no noise clustering algorithm for example 2

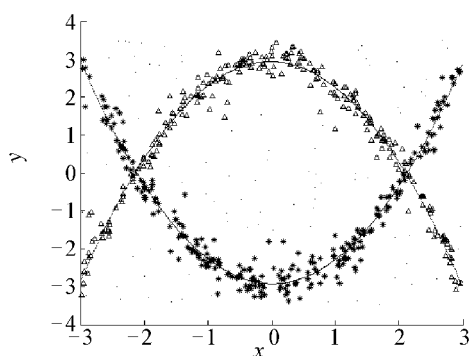


图7 实例2 NRC算法聚类结果

Fig. 7 Clustering results of noise resistance clustering algorithm for example 2

表1 无噪音算法和NRC算法得到的结果比较

Table 1 Results comparison of no noise clustering algorithm and NRC algorithm

	β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}
无噪音聚类算法	2.734 6	-0.043 7	-0.597 2	-2.706 7	-0.010 6	0.619 4
NRC算法	2.933 5	-0.022 8	-0.652 1	-2.927 7	-0.011 3	0.650 2

通过实验结果可以看出,NRC算法明显削弱了噪音数据的影响,得到了比无噪音聚类算法更优的模型参数.实验中 w 的选择也是很重要的, w 选择的太大,对噪音的抵抗能力就会减弱, w 选择的太小,就会将一些非噪音数据误当作噪音数据.

5 结束语

首先介绍了切换回归模型的基本概念,然后介绍了已有的切换回归模型的聚类方法,分析并指出了它们的缺陷:即它们都没有考虑噪音数据对聚类结果的影响.在此基础上提出了具有抵抗噪音能力的NRC算法,并通过仿真实验验证了NRC算法的优越性.实验上来看,NRC算法是相当成功的,但是仍然有一些工作需要做更深入的研究.模型中的子模型个数如何确定的问题,距离阈值 w 的选择以及算法的鲁棒性问题都将是未来研究的重点.

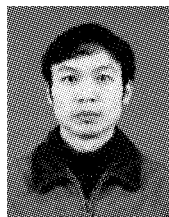
参考文献:

- [1] HAN Jiawei, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京:机械工业出版社, 2001: 223-261.
- [2] HAMERMESH D S. Wage bargains, threshold effects, and the Phillips curve[J]. Quarterly Journal of Economics, 1970, 84(3): 501-517.
- [3] QUANDT R E. A new approach to estimating switching regressions[J]. J Amer Statist Ass, 1972, 67(338): 306-310.
- [4] QUANDT R E, RAMSEY J B. Estimating mixtures of normal distributions and switching regressions[J]. J Amer

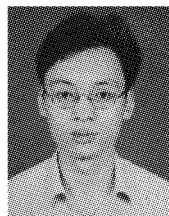
Statist Ass, 1978, 73: 730-752.

- [5] HOSMER D W. Maximum likelihood estimates of the parameters of a mixture of two regression lines[J]. Communications in Statistics, 1974, 3(10): 995-1005.
- [6] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum Press, 1981: 88-94.
- [7] HATHAWAY R J, BEZDEK J C. Switching regression models and fuzzy clustering[J]. IEEE Trans on Fuzzy Systems, 1993, 1(3): 195-204.
- [8] OHTA T, YAMAKAWA A, ICHIHASHI H, et al. Projection pursuit switching regression[C]//Proc of 5th International Conference on Soft Computing. Iizuka, Japan, 1998: 775-778.
- [9] OHTA T, YAMAKAWA A, ICHIHASHI H, et al. Projection pursuit switching regression for analysis of psychological feelings[J]. Journal of Biomedical Soft Computing and Human Sciences, 1998, 4(1): 15-21.
- [10] 沈红斌, 王士同, 吴小俊. 离群模糊核聚类算法[J]. 软件学报, 2004, 15(7): 1021-1029.
SHEN Hongbin, WANG Shitong, WU Xiaojun. Fuzzy kernel clustering with outliers[J]. Journal of Software, 2004, 15(7): 1021-1029.
- [11] WANG Shitong, JIANG Haifeng, LU Hongjun. A new integrated clustering algorithm GFC and switching regression[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2002, 16(4): 433-446.
- [12] 陆宏钧, 江海峰, 王士同. 关于切换回归的集成模糊聚类算法 GFC[J]. 软件学报, 2002, 13(10): 1905-1914.
LU Hongjun, JIANG Haifeng, WANG Shitong. An integrated fuzzy clustering algorithm GFC for switching regressions[J]. Journal of Software, 2002, 13(10): 1905-1914.

作者简介:



杨小兵,男,1976年生,博士,副教授,硕士生导师,主要研究方向为数据挖掘、知识工程等,发表学术论文10余篇,其中多篇被SCI、EI检索。



何灵敏,男,1974年生,博士,副教授,硕士生导师,主要研究方向为数据挖掘、机器学习等。



孔繁胜,男,1946年生,教授,博士生导师,主要研究方向为知识工程、数据挖掘、人工智能等,获国家科技进步三等奖1项、省部级科技进步一等奖1项、二等奖和三等奖各2项,1993年起享受国务院特殊津贴.在国内外重要刊物上发表学术论文30余篇,出版专著3部。