

特征加权和优化划分的模糊 C 均值聚类算法

肖林云,陈秀宏,林喜兰

(江南大学 数字媒体学院, 江苏 无锡 214122)

摘 要: 在传统 FCM 算法基础上提出一种称为特征加权和优化划分的模糊 C 均值算法(WGIFP-FCM). 该算法结合了优化划分法和特征权重思想,得到特征权重向量迭代公式并在聚类处理过程中自适应地更新. 仿真实验结果表明,该算法能科学地计算出样本特征权重向量,获得优于 GIFP-FCM、WFCM 和 FCM 的聚类效果,具有一定实际应用价值.

关键词: 聚类;模糊聚类算法;特征权重;聚类精度

中图分类号: TP274

文献标识码: A

文章编号: 1000—7180(2016)10—0143—04

Feature Weighted and Improved Partition Fuzzy C-Means Cluster Algorithm

XIAO Lin-yun, CHEN Xiu-hong, LIN Xi-lan

(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract: This paper proposes a new algorithm based on the FCM, which is called feature weighting and optimal fuzzy partition of fuzzy c-means algorithm (WGIFP - FCM). This algorithm combines the optimization partition method and the feature weight idea, obtains feature weight vector formula and updates it in the entire clustering process. Simulation experiments have proved that the proposed algorithm has gotten better clustering results than GIFP - FCM, WFCM and FCM while calculating the weighting vector scientifically, and so it will be available in practice.

Key words: clustering Algorithm; Fuzzy C-Means; feature weight; clustering accuracy

1 引言

作为数据处理的一种重要技术,聚类已经被广泛应用到诸如模式识别、图像处理和数据挖掘等许多领域^[1]. 聚类算法能找出隐藏在无标签数据集中的结构,并按照某种相似性度量使具有相同性质的数据尽可能归于同一类,它一般可分层次法、划分法、密度法和网格法等几类方法. 这些方法中,基于隶属度矩阵的聚类方法得到了广泛地的使用.

划分法又可以分为模糊聚类和可能性聚类两种,其中模糊 C 均值算法(FCM)^[2]是最常用的模糊聚类算法之一. 大多数最新提出的模糊聚类算法都是在它的基础上做的改进.

许多时候,样本的某个特征对整个数据集的聚类效果起着关键的作用. 然而,大多数 FCM 的改进算法并没有观察到这一点,这也是导致许多改进算法聚类精度不高、实用性不强的原因. 针对该问题,本文从样本的特征属性出发,考虑到不同样本特征对聚类的影响程度不同,为每一个样本特征引入相应的权重,同时结合优化划分法得到一种特征加权和优化 FCM 算法(WGIFP-FCM).

本文分析了常见的 FCM 算法,给出了其目标函数,对模糊划分矩阵和聚类中心迭代公式进行了推导. 在 FCM 算法的基础上结合优化划分法和特征权重的思想,依据拉格朗日求极值法求得隶属度、聚类中心和特征权重向量迭代公式,分析了改进算

法的实现过程,并将其应用到数据集聚类实验中.仿真结果验证了所提出算法的有效性.

2 传统模糊 C 均值(FCM)聚类算法介绍

模糊 C 均值(FCM)聚类算法.对给定的数据集 $\mathbf{X} = \{x_1, x_2, \dots, x_n\} \subseteq R^d$, 记隶属度矩阵为 $\mathbf{U} = (u_{ij})_{n \times c}$, $\mathbf{V} = [v_1, v_2, \dots, v_c]$ 为由 c 个类中心点组成的矩阵,则在 FCM 算法中类内平方误差函数 $J(\mathbf{X}, \mathbf{U}, \mathbf{V})$ 为

$$J(\mathbf{X}, \mathbf{U}, \mathbf{V}) = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m d^2(x_i, v_j) \quad (1)$$

式中, $m \in [1, +\infty)$ 是模糊指数,它控制着整个算法的模糊程度,对于不同的数据集, m 的值应该做相应的调整; c 是该数据集聚类的类别数且 $2 \leq c < n$; $u_{ij} (1 \leq i \leq n, 1 \leq j \leq c)$ 表示 x_i 在第 j 个类的隶属度值,并满足以下三个约束条件:

$$\sum_{j=1}^c u_{ij} = 1, 1 \leq i \leq n \quad (2)$$

$$1 \geq u_{ij} \geq 0, 1 \leq i \leq n, 1 \leq j \leq c \quad (3)$$

$$n > \sum_{i=1}^n u_{ij} > 0, 1 \leq j \leq c \quad (4)$$

式中, v_j 表示第 j 类的类中心; $d^2(x_i, v_j)$ 表示样本 x_i 和类中心 v_j 的相似性度量,一般用欧氏距离来表示.

对问题(1) 利用拉格朗日乘法得:

$$v_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m}, 1 \leq j \leq c \quad (5)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}}, 1 \leq j \leq c, 1 \leq i \leq n \quad (6)$$

于是,有如下 FCM 聚类算法:

(1) 选定聚类数目 c 、模糊指数 m 、最大迭代次数 T 和阈值 ϵ ,并令迭代计数器 $t = 0$,初始化隶属度矩阵 $\mathbf{U}$;

(2) 利用式(5) 计算聚类中心;

(3) 利用式(6) 计算隶属度矩阵;

(4) 根据当前 $\mathbf{V}$ 和 $\mathbf{U}$, 计算目标函数(1) 的值,若迭代次数大于 T 或者相邻两次目标函数值差的绝对值小于阈值 ϵ , 则算法停止;否则,令 $t = t + 1$,转步骤(2).

3 优化划分的 FCM 算法

基于竞争学习^[3],有赢家通吃原则和赢家获得

更多原则两种竞争学习规则^[4],在 FCM 算法的基础上给出如下优化模糊划分策略:对于每个输入数据,令其与对应的聚类中心具有较高的隶属度,而相对于其他中心点的隶属度则尽可能地小.这样可提高聚类过程中的抗干扰能力,加快收敛速度.称之为 GIFP-FCM 算法,相应的目标函数为

$$J_{\text{GIFP-FCM}} = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m d^2(x_i, v_j) + \sum_{i=1}^n a_i \sum_{j=1}^c u_{ij} (1 - u_{ij}^{m-1}) \quad (7)$$

其中

$$a_i = a * \min\{d^2(x_i, v_s) | s \in \{1, \dots, c\}\} \quad (8)$$

且 $1 \leq a < 1$ 为参数.

由式(7) 中的第二项可知,为使目标函数达到最小,只要隶属度趋向于 0 或 1 即可,这样实现优化模糊划分的思想.利用拉格朗日乘法可得如下公式:

$$v_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m} \quad (9)$$

$$u_{ij} =$$

$$\frac{1}{\sum_{k=1}^c \left(\frac{d^2(x_i, v_j) - a * \min_{1 \leq s \leq c} d^2(x_i, v_s)}{d^2(x_i, v_k) - a * \min_{1 \leq s \leq c} d^2(x_i, v_s)} \right)^{\frac{1}{m-1}}} \quad (10)$$

GIFP-FCM 算法过程类似于 FCM,而当 $a = 0$ 时 GIFP-FCM 退化为 FCM.

4 特征加权和优化划分 FCM 算法介绍

GIFP-FCM 虽然加快了算法收敛速度,但忽略了样本特征权重对聚类的影响,导致算法的聚类精度不高.针对 GIFP-FCM 聚类精度低,算法的实用性不强的问题,本文从样本特征入手引入特征权重向量来标记样本的每个特征,提出了新的 FCM 算法——WGIFP-FCM.

许多 FCM 的改进算法^[1,5] 都是建立在样本特征同等重要的假设条件下.在实际生活中,不同的特征对聚类的影响程度显然是不同的,所以如何科学合理地确定样本特征权重对聚类来说尤为关键.已有的一些研究成果大多是在聚类处理前预先计算出样本特征权重向量,但这种权重向量不能实现在聚类处理过程中自适应地调整,且不能由目标函数得其迭代公式,算出来的特征权重向量很多时候并不符合实际情况,反而影响了整个聚类过程.针对上述问题,本文在聚类处理的过程中设置特征权重向量,

通过不断地更新以接近实际值,并创造性地结合优化划分方法提出了一个新的模糊聚类算法——WGIFP-FCM算法,相应的目标函数可定义为

$$J(\mathbf{U}, \mathbf{V}, \mathbf{w}; \mathbf{D}) = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m [d_{ij}^{(w)}]^2 + \sum_{i=1}^m a_i \sum_{j=1}^c u_{ij} (1 - u_{ij}^{m-1}) \quad (11)$$

式中, $d_{ij}^{(w)} = \|\text{diag}(\mathbf{w})(\mathbf{x}_i - \mathbf{v}_j)\|$, $\text{diag}(\mathbf{w})$ 表示特征属性权重向量 $\mathbf{w} = (w_1, w_2, \dots, w_d)^T$ 的对角化, a_i

如式(8)所定义. 隶属度矩阵满足 $\sum_{j=1}^c u_{ij} = 1$; 同样特征权重向量也满足 $\sum_{q=1}^d w_q = 1$. 与式(11) 相应的拉格朗日函数为

$$L = J(\mathbf{U}, \mathbf{V}, \mathbf{w}; \mathbf{D}) - \sum_{i=1}^n \lambda_i (\sum_{j=1}^c u_{ij} - 1) - \beta (\sum_{q=1}^d w_q - 1) \quad (12)$$

分别关于 $\mathbf{U}, \mathbf{V}, \mathbf{w}$ 求偏导数得到以下公式:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{[d^{(w)}(\mathbf{x}_i, \mathbf{v}_j)]^2 - a * \min_{1 \leq s \leq c} [d^{(w)}(\mathbf{x}_i, \mathbf{v}_s)]^2}{[d^{(w)}(\mathbf{x}_i, \mathbf{v}_k)]^2 - a * \min_{1 \leq s \leq c} [d^{(w)}(\mathbf{x}_i, \mathbf{v}_s)]^2} \right)^{\frac{1}{m-1}}} \quad (13)$$

$$v_j = \frac{\sum_{i=1}^n u_{ij}^m \mathbf{x}_i}{\sum_{i=1}^n u_{ij}^m} \quad (14)$$

$$w_q = \frac{1}{\sum_{l=1}^d \left[\frac{\sum_{i=1}^n \sum_{j=1}^c u_{ij}^m (\mathbf{x}_{iq} - \mathbf{v}_{jq})^2}{\sum_{i=1}^n \sum_{j=1}^c u_{ij}^m (\mathbf{x}_{il} - \mathbf{v}_{jl})^2} \right]} \quad (15)$$

于是,可得以下 WGIFP-FCM 算法:

(1) 选定聚类数目 c 、模糊指数 m 、最大迭代次数 T 和阈值 ϵ , 并令迭代计数器 $t = 0$, 初始化划分矩阵 $\mathbf{U}$, 初始化特征属性权重向量 $\mathbf{w}$ (一般令每个值都为 $1/d$, d 为样本特征数);

(2) 根据式(14) 计算新的聚类中心;

(3) 根据式(13) 计算新的隶属度矩阵;

(4) 同时根据式(15) 更新属性权重向量 $\mathbf{w}$;

(5) 根据当前的 $\mathbf{V}, \mathbf{U}$ 和 $\mathbf{w}$, 计算目标函数(11) 的值. 若迭代次数大于 T 或者当前目标函数值与上次目标函数值之差的绝对值小于阈值 ϵ , 则算法停止运行; 否则, 令 $t = t + 1$ 并跳转(2).

5 实验结果与分析

为了验证本文提出的 WGIFP-FCM 算法的性

能, 分别在人造数据集和 UCI 真实数据集^[6] 中进行对比仿真实验; 用 FCM、WFCM 和 GIFP-FCM 作为比较算法. 实验环境是 Intel Core (TM) 2 Duo CPU E7200 2.53 GHz; 2.00 GB 内存; Windows 7; Matlab 2012b. 用归一化互信息 (Normalized Mutual Information, NMI)^[7] 和 芮氏指标 (Rand Index, RI)^[8-9] 作为评价指标:

(1) RI 指标

$$RI = \frac{a+b}{n(n-1)/2} \quad (16)$$

式中, a 表示具有不同的类标签且属于不同类的数据个数; b 表示具有相同的类标签且属于同一类的数据个数; n 表示样本的容量大小.

(2) NMI 指标

$$NMI = \frac{\sum_{i=1}^c \sum_{j=1}^c N_{i,j} \log \frac{N \times N_{i,j}}{N_i \times N_j}}{\sqrt{(\sum_{i=1}^c N_i \log N_i / N) \times (\sum_{j=1}^c N_j \log N_j / N)}} \quad (17)$$

式中, $N_{i,j}$ 表示第 i 个聚类与类 j 之间的契合度; N 表示样容量的大小; N_i 表示第 i 个聚类的样本数目; N_j 表示第 j 个聚类的样本数目.

一般地, $0 \leq NMI \leq 1, 0 \leq RI \leq 1$ 这两个指标的值越大, 说明该算法的聚类效果也就越好.

5.1 人造数据集实验

采用了两组高斯分布的模拟数据, 该数据集由高斯函数在设定类中心、类方差以及数据个数时随机生成. 图 1(a) 所示的人造数据集 1 分为三类, 每类均有符合高斯分布的 100 个数据样本, 三个类的聚类中心分别为 (3, 2), (3, 4) 和 (3, 6), 类方差矩阵都为 $\begin{pmatrix} 2 & 0 \\ 0 & 0.2 \end{pmatrix}$, 它表示围绕类中心随机分布的数据

点的波动范围, 该数据集样本的水平特征对聚类的影响远小于垂直特征. 图 1(b) 所示的人造数据集 2 分为三类, 每类均有符合高斯分布的 400 个数据样本. 三个类的聚类中心分别为 (2, 2), (4, 4) 和 (6, 6),

类方差矩阵都为 $\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, 该数据集样本的水平

特征对聚类的影响等于垂直特征. 为了保证实验结果的可比较性, 统一初始化迭代阈值 $\epsilon = 1e-5$, 最大迭代次数为 800, 模糊指数 m 值为 2, 聚类算法的实验结果如图 2 和图 3 所示.

由图 2 知, 对于人造数据集 1 来说 WGIFP-FCM 算法的聚类结果相比 FCM、WFCM 和 GIFP-

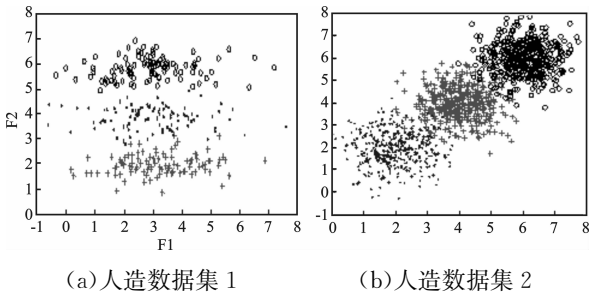


图 1 人造数据集

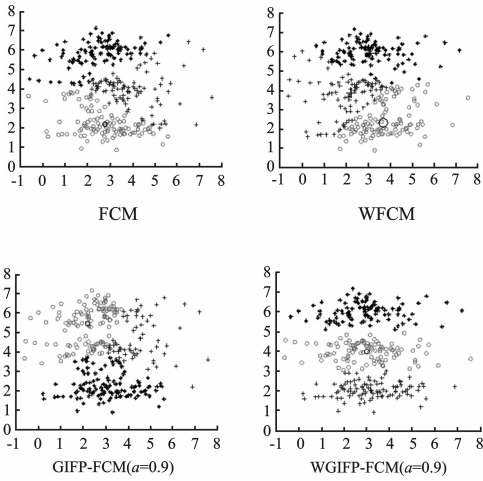


图 2 人造数据集 1 聚类结果

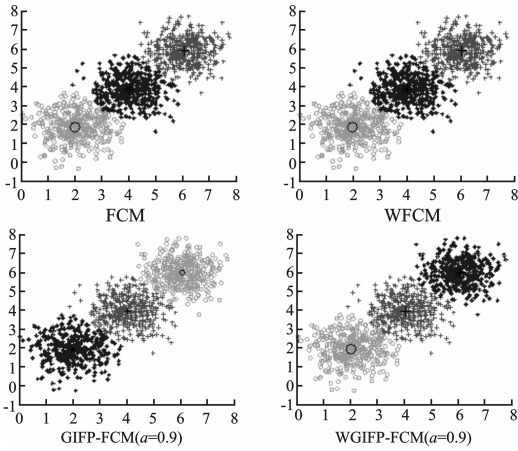


图 3 人造数据集 2 聚类结果

FCM 算法更加接近原数据的分布. WGIFP-FCM 算法的收敛速度较 FCM 算法大约提升了六倍,较 WFCM 算法提升了四倍左右,较 GIFP-FCM 算法提升了三倍左右;聚类精度高达 93.63% 以上,较前面几种算法提高了 10%~20%. 而且由 w 向量值可以看到, WGIFP-FCM 算法计算出的权重向量相比 WFCM 算法的计算结果更加接近实际情况(样本水平特征权重值远小于垂直特征权重值).

图 3 是四种算法在人造数据集 2 上的聚类结果分布图,从实验结果可以看出,对于一般数据集而

言, WGIFP-FCM 算法同样有着和其他算法相当的聚类性能表现.

5.2 真实数据集实验

为了验证在高维数据集中 WGIFP-FCM 算法的聚类性能,采用 UCI 真实数据库中的 wine 数据集和 segment 这两个数据集,它们的基本信息如表 1 所示. 实验中,统一初始化迭代阈值 $\epsilon=1e-5$,最大迭代次数为 800,模糊指数 m 值为 1.7.

表 1 实验中所使用的 UCI 数据集

Dataset Name	Instances	Attributes	Clusters
Segment	2 310	18	7
Wine	178	13	3

实验结果给出了 FCM、GIFP-FCM、WFCM 和 WGIFP-FCM 算法的聚类效果. WFCM 算法和 FCM 算法的收敛速度和聚类精度相差无几. 随着参数 a 值的增加, GIFP-FCM 和 WGIFP-FCM 的收敛速度不断提升, GIFP-FCM 的聚类精度较原始 FCM 算法只有小幅度的提高,而 WGIFP-FCM 却有大幅度的提高. 在同样 a 参数值的情况下, WGIFP-FCM 不论是在收敛速度还是聚类精度上都较 GIFP-FCM 算法有大幅度地提高. 真实数据集仿真实验也说明了 WGIFP-FCM 算法的聚类效果优于 FCM、WFCM 和 GIFP-FCM.

6 结束语

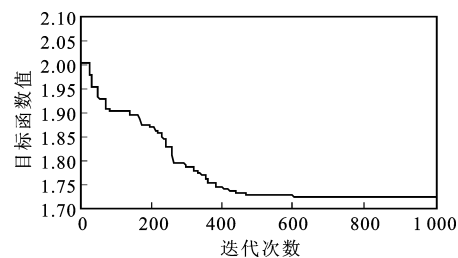
本文在传统 FCM 基础上结合优化模糊划分法和样本特征加权的思想,提出了一种新的 FCM 改进算法 WGIFP-FCM. 该算法在优化模糊划分的同时能由目标函数推导出特征权重向量迭代公式,比较合理地计算出样本的特征权重向量. 人造数据集和 UCI 真实数据集的仿真实验都表明, WGIFP-FCM 算法不仅解决了 FCM 算法的收敛速度慢问题,而且解决了聚类精度低的问题,验证了所提算法的有效性.

参考文献:

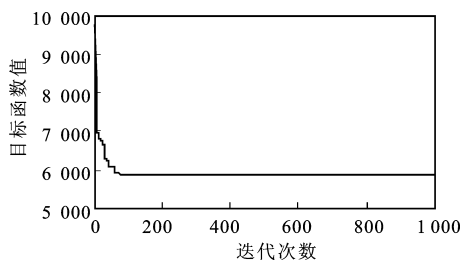
[1] 李朝锋, 居红云, 王琪. 基于 QPSO 的模糊 C 均值聚类算法[J]. 微电子学与计算机, 2008, 25(7): 194-197.

[2] Dunn J C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters[J]. Cybernetics, 1973, 3(3): 32-57.

[3] Lai Chung F, Lee T. Fuzzy competitive learning[J]. Neural Networks, 1994, 7(3): 539-551.



(a)传统方法



(b)本文方法

图3 网络接口跨平台移植扩展控件兼容性能对比

物理体系和通信体系分析,采用中间件技术进行网络节点结构的优化部署设计,然后采用线性规划调度算法进行网络接口跨平台移植的动态扩展资源调度设计,以此为基础实现对网络接口跨平台移植扩展控件设计方法的改进,仿真实验进行了性能测试,展示了本文设计方法在提高网络接口空间兼容性方面的优越性能,研究结果展示了较好的应用价值.

参考文献:

[1] 曾志,王晋,杜震洪,等. 一种云格环境下可计算资源与服务高效调配机制[J]. 浙江大学学报(理学版),2014,41(3):353-357.
[2] 杨照峰,王启明,吕海莲. 基于任务延迟的云计算资源

调度算法研究[J]. 计算机测量与控制,2014,22(2):499-502.
[3] 吕慧,赵跃华. 基于掩码的 AES 算法抗二阶 DPA 攻击方法研究[J]. 电子设计工程,2015,22(20):31-33.
[4] Badihi H,Zhang Y,Hong H. Model reference adaptive fault-tolerant control for a wind turbine against actuator faults[C]//2013 Conference on Control and Fault-Tolerant Systems (SysTol). Piscataway, NJ, USA: IEEE,2013:498-503.
[5] Chen L,Shi F,Patton R. Active FTC for hydraulic pitch system for an off-shore wind turbine[C]//IEEE Conference on Control and Fault-Tolerant Systems (SysTol). Piscataway,NJ,USA;IEEE,2013:510-515.
[6] Bakka T,Karimi H R,Christiansen S. Linear parameter-varying modeling and control of an offshore wind turbine with constrained information[J]. IET Control Theory&Applications,2014,8(1):22-29.
[7] 宋锦,吴定会,李意扬,等. 基于虚拟传感器的风力机 LPV 容错控制[J]. 信息与控制,2015,44(4):474-480,486.
[8] 卢宝莉,刘育梁,孙亮. 基于小尺度测量的双目立体视觉系统误差分析[J]. 光子学报,2015,44(10):1011001.
[9] 吴常清,王慧敏,薛涛. 基于 CloudStack 的私有云平台的构建与实现[J]. 西安工程大学学报,2014,28(2):220-224.

作者简介:

魏惠茹 女,(1982-),硕士,讲师. 研究方向为网络工程、数据挖掘. E-mail:weihuiru@126.com
李晓东 男,(1980-),硕士,讲师. 研究方向为软件工程,面向对象.
朱齐亮 男,(1982-),硕士,讲师. 研究方向为软件工程,计算机应用.

(上接第 146 页)

[4] Zhu L, Chung F L, Wang S. Generalized fuzzy c-means clustering algorithm with improved fuzzy partitions[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(3): 578-591.
[5] 王洪春,彭宏. 基于模糊 C-均值的增量式聚类算法[J]. 微电子学与计算机,2007,24(6):156-157.
[6] Pal N R, Bezdek J C. On cluster validity for the fuzzy c-means model[J]. IEEE Transactions on Fuzzy Systems, 1995, 3(3): 370-379.
[7] Ghosh J. Multiclassifier systems: Back to the future [M]. BerlinHeidelberg:Springer, 2002: 1-15.
[8] Iwayama M, Tokunaga T. Hierarchical bayesian clus-

tering for automatic text classification[C]// Proceedings of the 14th international joint conference on Artificial intelligence. San Francisco, CA, USA: ACM, 1995: 1322-1327.
[9] 朱红霞,王黎明. 改进的自适应模糊 C 均值聚类图像分割算法[J]. 微电子学与计算机,2010,27(12):87-89.

作者简介:

肖林云 男,(1988-),硕士研究生. 研究方向为模式识别、图像处理. E-mail:18206181357@163.com.
陈秀宏 男,(1964-),博士后,教授. 研究方向为模式识别、图像处理、人工智能.
林喜兰 女,(1989-),硕士研究生. 研究方向为数字图像处理.