

聚类分析在肺结节识别中的应用

孙娟¹, 王兵^{1*}, 杨颖², 田学东¹

(1. 河北大学 数学与计算机学院, 河北 保定 071002; 2. 河北大学附属医院 CT 室, 河北 保定 071000)

(* 通信作者电子邮箱 wangbing@hbu.edu.cn)

摘要:针对肺部微小结节难于识别的问题,提出用聚类算法分析肺部感兴趣区域(ROI)的方法。为进一步提高运行速度和识别率,提出全权模糊聚类算法 PWFCM,给每个样本及其特征分别赋予权值并引入新的隶属度约束改进收敛性;利用二次聚类策略降低不均衡 ROI 数据造成的低敏感度。对实际 CT 影像数据进行测试,实验结果表明:该聚类分析具有高敏感度和低假阳性率,能有效地检测出肺结节。

关键词:医学图像;计算机断层扫描;感兴趣区域;肺结节;模糊 C 均值聚类

中图分类号: TP393 **文献标志码:** A

Research on cluster analysis in pulmonary nodule recognition

SUN Juan¹, WANG Bing^{1*}, YANG Ying², TIAN Xuedong¹

(1. College of Mathematics and Computer Science, Hebei University, Baoding Hebei 071002, China;

2. CT Department, Hebei University Affiliated Hospital, Baoding Hebei 071000, China)

Abstract: Aiming at the problem of pulmonary small nodules was difficult to identify, a method using fuzzy C-means clustering algorithm to analyse the lung Region Of Interest (ROI) was presented. An improved Fuzzy C-Means clustering algorithm based on Plurality of Weight (PWFCM) was presented to enhance the accurate rate and speed of small nodules recognition. To improve the convergence, each sample and its features were weighted and a new membership constraint was introduced. The low sensitivity from the uneven ROI data was decreased by using a double clustering strategy. The experimental results tested on the real CT image data show that PWFCM algorithm can detect lung nodules with a higher sensitivity and lower false positive rate.

Key words: medical image; Computed Tomography (CT); Region Of Interest (ROI); pulmonary nodule; fuzzy C-means clustering

0 引言

肺癌已经被公认为是威胁人类生命的主要疾病。肺癌的早期诊断是有效挽救生命的关键手段,而肺癌的早期并没有什么明显病症,多数是通过体检的计算机断层扫描(Computed Tomography, CT)图像以肺结节的形式表现。肺结节特别是微小结节在早期影像诊断中由于体积小、对比度低等原因很容易被忽略^[1]。而一个病人的高分辨率扫描图片多达数百张,在阅片时很容易误判、漏判。因此,迫切需要计算机辅助诊断技术把这些复杂图像信息及相互关系以直观的方式显示给医生,提高诊断的准确性和科学性。

肺结节通常指直径小于 3 cm 的类圆形病灶,其形状各异,易与肺内血管、气管混淆,且分布位置不定,容易和肺内其他组织粘连。肺部感兴趣区域(Region Of Interest, ROI)通常指在 CT 影像中疑似肺结节点。肺结节检测技术经过对原始 CT 影像的图像处理,进行肺实质分割——ROI 分割,ROI 特征提取最后进行分类找出肺结节。目前多数算法在初始分割 ROI 后,使用分类方法(如支持向量机(Support Vector

Machine, SVM)、神经网络等)对 ROI 数据分类找出肺结节^[2-7]。分类算法是有导师学习,需要标记有类信息的原始 ROI 数据作为训练数据。但由于现实问题,收集到的原始数据信息量少且不全面,使用有导师的分类算法很难产生有效的分类器。

聚类分析为无导师学习,通过运用某种相似性的度量方法,将相似度大的模式尽量聚为一类(簇)。目前在肺结节智能识别系统中,聚类方法多用于肺部 ROI 区域的初始分割^[8-9],很少对 ROI 数据聚类发现肺结节。只有文献[10]使用 3 种聚类方法直接从 ROI 区域分割出肺结节,但准确率较低。文献[11]使用聚类算法分别求出肺结节和非肺结节的聚类中心,根据欧氏距离对未知样例进行分类。该方法属于分类且敏感度低,分类效果不好。而聚类算法作为无导师学习方法不需要类别信息进行训练,可以避免由于原始数据量不足存在的人为噪声。

本文在充分考虑肺结节检测中肺结节与非肺结节(肺内血管、气管)数量的不均衡性后,结合肺部 ROI 数据的特征,提出了全权模糊聚类算法(Fuzzy C-Means algorithm based on

收稿日期:2014-01-20;修回日期:2014-03-07。

基金项目:国家自然科学基金资助项目(10804025, 61375075);河北省自然科学基金资助项目(F2012201020, F2014201098)。

作者简介:孙娟(1975-),女,河北保定人,讲师,硕士,主要研究方向:机器学习、图像处理;王兵(1966-),女,河北承德人,教授,主要研究方向:图像处理模式识别;杨颖(1970-),女,河北保定人,硕士,主要研究方向:医学影像;田学东(1963-),男,河北保定人,教授,博士,主要研究方向:模式识别与图像处理、中文信息处理。

Plurality of Weight, PWFCM)。根据肺结节的临床病理特征,不但为每个实例的重要属性赋予权值,还为每个样本赋予权值,再引入改进后的隶属度函数提高聚类结果的收敛速度,且应用二次聚类策略降低数据的不均衡性。观测肺门附近的CT影像,如图1(a)医生很容易发现直径较大的结节,但如图1(b)直径较小的小结节很容易与血管气管混淆被视为阴性。本文重点查找在肺部直径小于1 cm的微小结节。

实验测试使用二次聚类策略:第一次聚类去除面积较大的ROI,主要包括面积较大扁长形状的肺内血管、气管;第二次在数量趋于均衡、形状类圆的ROI中聚类找到阳性结节。实验结果表明该算法能快速、有效地标识出疑似肺结节。

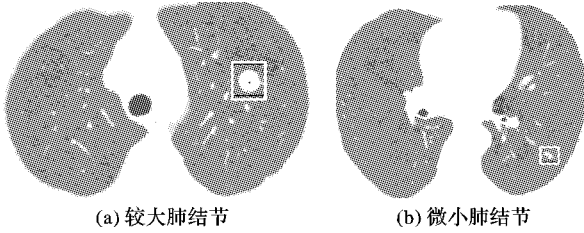


图1 同一个病人不同位置的两张CT影像

1 基本概念

1.1 肺结节智能识别系统

肺结节智能识别是计算机辅助诊断的重要部分,在原始CT影像上需要先分割出肺实质、候选肺结节分割、肺结节特征选择与提取、肺结节分类识别等。图2给出了典型的肺结节检测示意图。

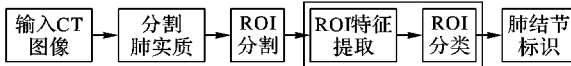


图2 肺结节检测流程

在整个检测识别过程中,ROI的特征提取和分类判定是最重要的部分,影响整个检测系统的性能。

1.2 ROI特征选取

肺部ROI特征选取直接影响聚类结果。合适的特征选取能够避免聚类时产生混乱,从而使聚类结果不准确。

多数研究中,ROI的特征提取选用的特征有:形状特征参数如面积、周长、圆形度等特征;CT特征参数如灰度均值、方差等和傅里叶描述子等。魏颖等^[3]提出使用紧凑度和低阶不变矩描述子等11个特征应用SVM分类器进行肺结节检测。

但过多的选择特征会降低分类准确率,合理的特征组合必须满足以下三个方面要求:1)这些特征可以很好地描述肺部ROI的各种特征;2)具有几何变换不变性;3)能够明显地区分出结节与非结节。本文从ROI的形态特征、灰度特征和纹理特征中选用了6个特征。

1)区域面积(S):ROI区域的总体像素个数。

2)圆形度($Circularity$):反映ROI接近圆形的程度,在圆形边界时取最小值。 $Circularity = 4\pi S/L^2$ 。其中周长 L 为ROI边缘像素点个数。

3)似圆度(N):求出区域周长上距离最远的两个像素点,以它们距离为直径做外接圆,其面积为 S' ,似圆度: $N = S'/S$ 。

4)边界离心率($Eccentricity$): $Eccentricity = D/D'$,其中 D

为ROI直径, D' 为垂直于直径 D 的最长弦。

5)与肺门间的距离:计算肺门位置 M 与第 i 个ROI中心 m_i 的距离。 $Dis_i = \|M - m_i\|$ 。

6)平均灰度值: $mean = \frac{1}{N} \sum_{(i,j) \in ROI} g(i,j)$,其中 $g(i,j)$ 代表ROI像素点 (i,j) 的灰度值。

1.3 模糊C均值算法

模糊C均值聚类(Fuzzy C-Mean, FCM)算法是目前应用最广泛的聚类算法之一。FCM是一种使目标函数最小化的逐步迭代过程,使用取值范围在0与1之间的隶属度函数体现某些对象所具有的模糊特征。但FCM算法的聚类数目很难取定,并对初始聚类中心敏感,易于陷入局部极值。为此,国内外学者提出了多种C均值的改进算法,张靖等^[12]多次调用聚类算法,根据 k 个类中心的个体轮廓系数以及各样本与类中心的距离,自适应地选取优秀样本,求其均值作为初始聚类中心。但时间复杂度高。朱林等^[13]等引入了新的隶属度函数,聚类收敛速度很快,但由于使用了随机初始化隶属度矩阵,可能导致聚类结果收敛到一致解。张瑞丽等^[14]提出的W-距离均值FCM算法,为每个样本赋予权值,提高算法的抗噪性,但该算法在处理不平衡数据时存在缺陷。本文在文献[14]算法的基础上,提出全权模糊聚类算法,针对肺部ROI数据特点,对每个特征加入了权重并修改隶属度函数,提高算法收敛速度。

设样本集 $X = \{X_1, X_2, \dots, X_N\}$, N 为样本总数, X_i 是 n 维向量, $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$, n 为特征个数, x_{ij} 是样本 X_i 的第 j 个特征($j = 1, 2, \dots, n; i = 1, 2, \dots, N$)。 $U = \{u_{ij}\}$ 其中 u_{ij} 是样本数据 X_j 隶属于第 i 类的隶属度($j = 1, 2, \dots, N; i = 1, 2, \dots, c$, c 为设定类别数)。隶属度之和满足

$$\sum_{i=1}^c u_{ij} = 1, \forall j = 1, 2, \dots, N \quad (1)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c (d_{ij}/d_{kj})^{2/(m-1)}}; i = 1, 2, \dots, c, j = 1, 2, \dots, N \quad (2)$$

其中 $d_{ij} = \|X_j - v_i\|$ 为欧氏距离。 $m \in [1, \infty)$ 为一个加权指数, m 越大聚类的模糊性越强,通常 m 取2。给定 $V = \{v_1, v_2, \dots, v_c\}$ ($v_i \subset V, i = 1, 2, \dots, c$)是聚类中心。

$$v_i = \left(\sum_{j=1}^N u_{ij}^m X_j \right) / \left(\sum_{j=1}^N u_{ij}^m \right) \quad (3)$$

FCM的目标函数为

$$J(U, V, X) = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^m d^2(X_j, v_i) \quad (4)$$

FCM算法把 N 个样例分成为 c 个模糊簇,并求每簇的聚类中心 v_i 使得准则函数 J 达到最小。

2 全权模糊聚类算法

肺部ROI区域间的相似性是通过距离公式计算。考虑到每个特征的重要度不同,每个ROI区域对聚类结果的影响也是不同的,本文为每个样本点赋予样本权值 w_i ,为不同特征设定贡献度 W_k 。 $w_i = N[i]/N$ ($i = 1, 2, \dots, N$), $N[i]$ 是每个样本 i 小于样本距离阈值 d 所对应的个数, $N[i]$ 越大则其权值越大,说明样本点 i 必处于数据分布密集区的中心,对聚

类结果影响较大。 $W_k \in [0, 1] (k = 1, 2, \dots, n)$ 是与输入属性相对应的一个权重,描述第 k 维属性在聚类中的重要性。 $W_k = 1$ 时,第 k 维属性在聚类中发挥全部作用, W_k 的值越小,该属性在聚类过程中发挥的作用越小, W_k 为0时,该属性在聚类中不发挥任何作用。

PWFCM 算法重新定义了样本相似性度量函数,使用基于欧氏距离加权的相似性度量,定义样本 p 与样本 q 的距离为:

$$d_{pq}^w = \sqrt{\sum_{i=1}^n W_i^2 (x_{pi} - x_{qi})^2} \quad (5)$$

隶属度定义重新定义为:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d^{w^2}(X_j, v_i) - \alpha \times \min_{1 \leq s \leq c} d^{w^2}(X_j, v_s)}{d^{w^2}(X_j, v_k) - \alpha \times \min_{1 \leq s \leq c} d^{w^2}(X_j, v_s)} \right)^{\frac{1}{m-1}}} \quad (6)$$

计算聚类中心和目标函数 J 的公式重新定义如下:聚类中心:

$$v_i = \frac{\sum_{j=1}^N w_j u_{ij}^m X_j}{\sum_{j=1}^N w_j u_{ij}^m}; i = 1, 2, \dots, c \quad (7)$$

$$J(U, V, X) = \sum_{i=1}^c \sum_{j=1}^N w_j u_{ij}^m (d_{ji}^w)^2 \quad (8)$$

全权模糊聚类算法描述如下:

1) 计算距离矩阵 D 和样本间的距离均值阈值 \bar{d} 。

$$D = \sum_{i=1}^N \sum_{j=1}^N d_{ij}^w; i = 1, 2, \dots, N, j = 1, 2, \dots, N \quad (9)$$

$$\bar{d} = \frac{2\omega \times \sum_{i=1}^{N-1} \sum_{j=i+1}^N d_{ij}^w}{(n \times (n-1))}; 0 < \omega \leq 1 \quad (10)$$

2) 对每个样本 i 统计满足 $d_{ij}^w \leq \bar{d}$ 的样本 j 的个数,并把得到的个数存入数组 $N[i]$ 中。从 $N[i]$ 中找出取值最大的一个作为首选初始聚类中心,从其余 $N[i]$ 中再找取值最大的,直到找到 c 个初始聚类中心。

3) 运用式(6)计算 X_i 在第 i 个聚类中的模糊隶属度。

4) 根据式(7)更新聚类中心 v_i 。

5) 根据式(8)计算目标函数,若对比上一次目标函数值的改变量小于给定的阈值,则此算法停止;否则转向步骤3)。

3 实验步骤与结果

从一幅CT图像中获取感兴趣区域,首先需要从原图中分割出肺实质,然后再分割ROI并去除小区域。图3给出从CT图像中获取ROI的示例图。图3(c)中可以明显看出血管、气管的比例要远远高于肺结节的个数。为实现肺部感兴趣区域不均衡数据的聚类,实验采用二次聚类策略降低数据的不均衡性,使聚类结果更准确。

第一次聚类 调整属性权值,将面积的权重设为最高,使用PWFCM聚类3个簇,保留平均面积最小的疑似结节簇。如图3(d),平均面积最小的类圆形ROI簇被保留下来,平均面积较大的血管簇和线形的气管簇被排除为阴性。

第二次聚类 调整属性权值,面积的权值设为最小,圆形度权值设为最高,对疑似结节簇再次使用PWFCM聚类2个

簇。如图3(e),获得阳性肺结节和阴性非结节两个簇。

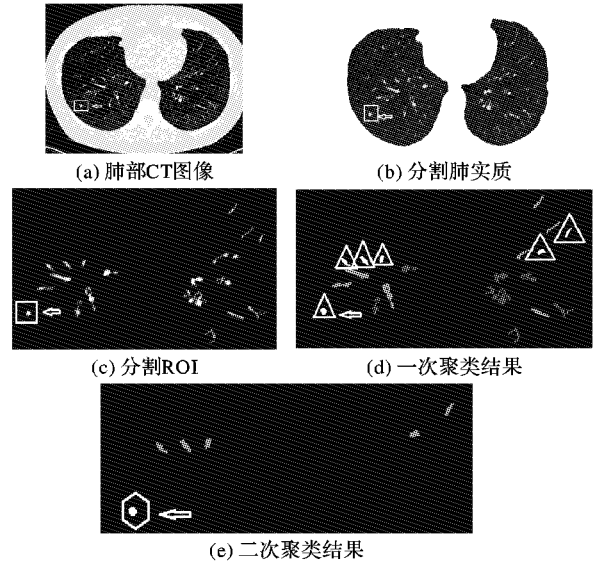


图3 二次聚类策略测试CT图像步骤

使用来自河北某大型医院放射科实际临床影像,选取30幅图像,使用二次聚类策略的全权模糊聚类算法进行聚类分析。肺结节位置不定,其中,每幅图有一个结节被医生标志。

使用敏感性和假阳性率两个指标分析基于二次聚类策略的PWFCM算法性能。敏感性定义为:

敏感性(sensitivity):

$$Sn = \frac{TP}{TP + FN} \quad (11)$$

假阳性率(false positive rate):

$$FPR = \frac{FP}{TN + FP} \quad (12)$$

其中: TP 为检测的真阳性肺结节, FN 为漏检真阳性肺结节, FP 为错判为阳性的假阳性非肺结节, TN 为真阴性非肺结节。30幅图像检测数据的平均敏感性为0.982,假阳性率为0.088。敏感性高于同类文献[3,8,11]中的数值。

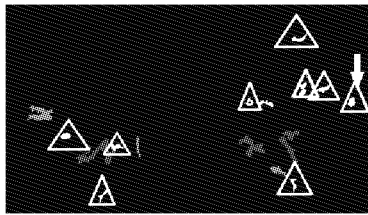
从聚类效果看,PWFCM比传统FCM算法的聚类更准确,如图4分别使用两种算法对同一CT影像分割后的ROI区域进行聚类,图4(b)的含肺结节的簇中的ROI形状更相似,而图4(a)中的一些线形小面积ROI也分入含肺结节的簇中。

对基于二次聚类策略的传统FCM算法进行实验测试,使用相同的30幅CT图像进行检测,算法的平均敏感性为0.946,假阳性率为0.194。

模糊聚类算法是一种迭代算法,先找到初始聚类中心,然后反复计算隶属度和聚类中心,直到算法满足收敛条件。算法的时间复杂度取决于样本个数 N ,样本特征维数 n ,聚类个数 c 和迭代次数 t 。FCM算法初始化聚类中心的时间复杂度是 $O(N \cdot \log N \cdot n)$ 而PWFCM算法是 $O(N^2 \cdot n)$;在迭代计算隶属度和聚类中心时,两种算法的时间复杂度均为 $O(t \cdot c \cdot N \cdot n)$ 。因为每张肺部CT影像的ROI样本的数量 N 很小(通常 $N < 30$),所以在实际测试时,两种算法初始化聚类中心所用时间几乎相同,但PWFCM算法找到的初始聚类中心更合理有效,易满足收敛条件,收敛速度快。对30幅肺部CT影像进行测试,WPFCM算法的平均迭代次数为46,运行时间为

3.76 ms;FCM 算法的平均迭代次数为 57,运行时间为4.1 ms。

实验结果表明,传统 FCM 算法的聚类速度要略低于 PWFCM 算法,敏感性低于 PWFCM 算法而假阳性率高于 PWFCM 算法。本文提出的 PWFCM 算法性能优于传统 FCM 算法。结合二次聚类策略的 PWFCM 算法可以很好地检测出体积小类圆形的肺结节。对于体积较大的肺结节,只使用 PWFCM 算法就能得到理想聚类效果。本方法对图像分割步骤的结果不敏感,不需过度精确分割 ROI,减少了分割 ROI 步骤中的过度图像分割导致的误删肺结节现象。原始 CT 影像只有灰度图像,而本文在聚类后着色的 ROI 簇可以辅助医生发现漏掉的微小肺结节。



(a) 一次FCM聚类结果



(b) 一次PWFCM聚类结果

图4 PWFCM 与 FCM 聚类结果比较(聚类中心为3)

4 结语

本文提出的全权模糊聚类方法为每个样本点和不同特征分别赋予权值,并自适应产生初始聚类中心,通过修改隶属度函数提高聚类收敛速度。实验表明全权模糊聚类算法比传统模糊聚类算法能更快和更准确地聚类不同性质 ROI 区域。二次聚类策略的提出,从根本上减少了肺部不平衡 ROI 数据对聚类结果的影响。实验结果表明,应用基于二次聚类策略的全权模糊聚类算法能有效地对 ROI 区域进行聚类,该方法具有较高敏感性和较低的假阳性率。为方便医生检测,用不同颜色标明相似 ROI 区域可直观提示疑似肺结节位置。

参考文献:

- [1] NIE S, SUN X, CHEN Z. Progress in computer-aided detection for pulmonary nodule using CT image[J]. Chinese Journal of Medical Physics, 2009, 26(2): 1075 - 1079. (聂生东, 孙希文, 陈兆学. 基于 CT 图像的肺结节计算机辅助检测技术的研究进展[J]. 中国医学物理学杂志, 2009, 26(2): 1075 - 1079.)
- [2] MICHELA A, MARCO C, BEATRICE L, et al. Computer-aided detection of lung nodules based on decision fusion techniques[J]. Industrial and Commercial Application, 2011, 14(1): 295 - 310.
- [3] WEI Y, GUO W, SUN Y, et al. Feature selection and classification algorithm for region of interest in lung cancer CAD system[J]. Information and Control, 2008, 12(4): 445 - 451. (魏颖, 郭薇, 孙月芳, 等. 面向肺癌 CAD 系统的感兴趣区域特征选择与分类算法[J]. 信息与控制, 2008, 12(4): 445 - 451.)
- [4] WEI Y, LI R, YANG J, et al. An algorithm for segmentation of lung ROI by mean-shift clustering combined with multi-scale HESSIAN matrix dot filtering[J]. Journal of Central South University, 2012, 19(12): 3500 - 3509.
- [5] GUO W, WEI Y, ZHOU H, et al. Adaptive detection algorithm for pulmonary nodules[J]. Journal of System Simulation, 2009, 21(13): 3955 - 3958. (郭薇, 魏颖, 周翰逊, 等. 肺结节的自适应检测算法[J]. 系统仿真学报, 2009, 21(13): 3955 - 3958.)
- [6] HE Z, LIANG Y, HUANG X, et al. Research on the feature extraction approach for SPNs detection[J]. Journal of Chinese Computer Systems, 2009, 30(10): 2074 - 2077. (何中市, 梁琰, 黄学全, 等. 肺结节检测中特征提取方法研究[J]. 小型微型计算机系统, 2009, 30(10): 2074 - 2077.)
- [7] WANG L, LI B, TIAN L, et al. Intelligent detection of lung nodules based on rules and multi-feature tracking[J]. Journal of Biomedical Engineering Research, 2010, 29(2): 79 - 83, 203. (王立非, 李彬, 田联房, 等. 基于规则及多特征跟踪的肺结节的智能检测方法[J]. 生物医学工程研究, 2010, 29(2): 79 - 83, 203.)
- [8] WU Y, YANG X, XU M. et al. Graph cuts medical image segmentation algorithm based on K-means clustering[J]. Computer Engineering, 2011, 37(5): 232 - 234. (吴永芳, 杨鑫, 徐敏, 等. 基于 K 均值聚类的图割医学图像分割算法[J]. 计算机工程, 2011, 37(5): 232 - 234.)
- [9] SUN X, TIAN Q, LI L, et al. Computer-aided detection algorithm for pulmonary nodule based on the improved fuzzy c-means cluster algorithm[J]. Progress in Modern Biomedicine, 2010, 10(17): 3326 - 3331. (孙旭辉, 田启川, 李临生, 等. 基于改进的模糊 C 均值聚类的肺结节计算机辅助诊断算法研究[J]. 现代生物医学进展, 2010, 10(17): 3326 - 3331.)
- [10] ALFONSO C, CARMEN B, ALBERTO R, et al. An analysis of different clustering algorithms for ROI detection in high resolutions CT lung images[C]// Proceedings of the 2010 International Conference on Computer Vision and Graphics. Berlin: Springer, 2010: 241 - 248.
- [11] CHEN K, LI B, TIAN L. Pulmonary nodules detection algorithm based on local threshold and iterative of clustering center[J]. Computer Science, 2012, 39(2): 302 - 304. (陈侃, 李彬, 田联房. 基于局部阈值和聚类中心迭代的肺结节检测算法[J]. 计算机科学, 2012, 39(2): 302 - 304.)
- [12] ZHANG J, DUAN F. Improved k-means algorithm with meliorated initial centers[J]. Computer Engineering and Design, 2013, 34(5): 1691 - 1694, 1699. (张靖, 段富. 优化初始聚类中心的改进 k-means 算法[J]. 计算机工程与设计, 2013, 34(5): 1691 - 1694, 1699.)
- [13] ZHU L, WANG S, DENG Z. Research on generalized fuzzy c-means clustering algorithm with improved fuzzy partitions[J]. Journal of Computer Research and Development, 2009, 46(5): 814 - 822. (朱林, 王士同, 邓赵红. 改进模糊划分的 FCM 聚类算法的一般化研究[J]. 计算机研究与发展, 2009, 46(5): 814 - 822.)
- [14] ZHANG R, ZHANG J. Fuzzy clustering algorithm based on w-mean distance[J]. Journal of Computer Applications, 2012, 32(7): 1978 - 1982, 1986. (张瑞丽, 张继福. 基于 w-距离均值的模糊聚类算法[J]. 计算机应用, 2012, 32(7): 1978 - 1982, 1986.)