文章编号：2095-2163(2020)06-0238-05 中图分类号：TP18 文献标志码：A

基于差分进化的加权 k-means 算法研究

王凤领
（贺州学院，广西 贺州 542899）

摘 要：针对处理大样本数据时聚类算法的局限性，以及 k-means 算法受初始聚类中心和异常数据的限制，聚类结果不稳定的问题，本文提出了基于差分进化的加权 k-means 算法，优先选择初始聚类中心，采用差分进化算法，根据样本对聚类分析影响程度不同，设计加权欧氏距离，来减少异常点带来的不利影响，从而获得稳定的聚类结果。实验结果表明，该算法选择的初始聚类中心更接近最终聚类中心，提高了算法的计算效率。

关键词：差分进化；k-means 算法；加权 k-means 算法

Research on a Weighted k-means Algorithm Based on Differential Evolution

WANG Fengling
（Hezhou University, Hezhou Guangxi 542899, China）

【Abstract】 In view of the limitations of clustering algorithm in dealing with large sample data, the limitation of k-means algorithm by initial clustering center and abnormal data, and the instability of clustering results, a weighted k-means algorithm based on differential evolution is proposed. The initial clustering center is selected first, and differential evolution algorithm is adopted. According to the different influence degree of samples on clustering analysis, the weighted euclidean distance is designed in order to reduce the adverse effects of abnormal points and obtain stable clustering results. The experimental results show that the initial cluster center selected by the algorithm is closer to the final cluster center, which improves the efficiency of the algorithm.

【Key words】 differential evolution; k-means algorithm; weighted k-means algorithm

0 引 言

差分进化算法是 Storn.R 和 Price.K 在 1995 年提出的基于种群进化的启发式算法，是随机并行全局搜索算法。它具有记忆个体最优解和在群体中共享信息的特点，即通过群体中个体之间的合作和竞争来解决优化问题。差分进化算法具有易于理解、简单高效、易于使用、鲁棒性好、收敛速度快、全局搜索能力强等优点。目前，差分进化算法已成功应用于许多研究领域，并取得了较好的效果。

1 差分进化算法

差分进化算法作为优化算法的一种，具有变异、交叉和选择操作。差分进化算法还具有控制参数少、收敛速度快、原理简单、全局优化能力突出等优点，其独特的差分变异操作可以保证种群的多样性，并能使种群向更好的方向发展。

1.1 差分进化算法的特点

(1)通用性。无需问题的先验信息。可用实数编码问题的可行解，并且不依赖于问题的信息。

(2)易于实现，结构简单。

(3)控制较少的参数。

(4)局部搜索和全局搜索协同搜索。

(5)具有记忆能力，能够记住群体搜索过程中个体的最优解。

(6)易于与其他算法相结合，构建性能更好的混合算法。

1.2 差分进化算法的步骤和过程

1.2.1 步骤

种群中的每个个体都是求解问题的可行解，种群规模为 N，个体的适应度为函数，表示第 t 代种群的第 i 个个体，F 是缩放系数，t 是进化代数。表示交叉概率。差分进化算法的步骤描述为：

第一步 完成种群初始化，设置群体大小 N，缩放因子 $F \in [0,2]$，交叉概率 $C_R \in [0,1]$，进化代数 $t=0$，随机生成初始种群 $X(0) = \{X_1(0), X_2(0),\ldots,X_N(0)\}$，其中，任一个体 $X_i(0)$ 包含 D 个分量的向量，即 $X_i(0) = \{X_{i1}(0), X_{i2}(0),\ldots, X_{iD}(0)\}$；

第二步 计算出每个个体的适应度值 $f(X_i(t))$，评价种群中的每个个体；

第三步 根据式(1)完成种群个体 $X_i(t)$ 变异

基金项目：贺州学院教授科研启动基金资助项目(HZUJS201615)。

作者简介：王凤领(1976-)，男，硕士，教授，主要研究方向：数据挖掘、网络与智能信息技术。

收稿日期：2020-03-13

操作，得到变异个体 $V_i(t)$；

$$V_i(t) = X_{ri}(t) + F(X_{r2}(t) - X_{r3}(t). \quad (1)$$

第四步　根据式(2)完成交叉操作，得到中间测试个体 $U_i(t)$，$U_i(t)$ 是由 D 维分量组成，即 $U_i(t) = (u_{i1}(t), u_{i2}(t), \ldots, u_{iD}(t))$。

$$u_{ij}(t) = \begin{cases} v_{ij}(t) \ randj(0,1) \leqslant C_R \text{ or } j = r4; \\ v_{ij}(t) \ randj(0,1) > C_R \text{ or } j \neq r4. \end{cases} \quad (2)$$

第五步　对于最小化求解问题，适应度函数值小的个体进入下一代种群中继续进化。在选择操作过程中，采取贪心策略，进行最佳选择，通过比较当前进化个体 $X_i(t)$ 与相对应中间试验个体的适应度值。选择方式按照公式(3)进行

$$X_i(t+1) = \begin{cases} U_i(t) \text{ if } f(U_i(t)) \leqslant f(X_i(t)); \\ X_i(t) \text{ if } f(U_i(t)) > f(X_i(t)). \end{cases} \quad (3)$$

第六步　检验种群 $X(t+1)$ 中的个体，如果终止算法满足条件，则输出；否则 $t = t + 1$，转到步骤二。

1.2.2　算法流程

根据差分进化算法的描述，获得差分进化算法的如图 1 所示的流程图。

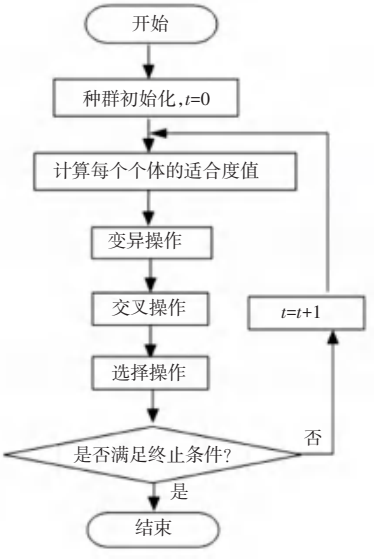

图 1　差分进化算法的流程图

Fig. 1　Flow chart of differential evolution algorithm

2　k-means 聚类算法

2.1　k-means 聚类算法概述

k-means 是 Hartigan 提出的一种基于划分的聚类方法，是一种基于距离划分的聚类算法。距离作为相似性评价标准。当两物体间的距离较近时，它们之间的距离就较小，其相似度较高。

2.2　k-means 算法的缺点

(1)容易陷入局部最优解；

(2)易受孤立点的影响；

(3)算法对初始中心选择具有随机性。

2.3　k-means 聚类算法步骤

k-means 聚类算法就是将 n 个数据样本对象分割成 k 个类别，把聚类个数 k 作为输入参数，使不同的类别间数据点的相似性尽可能小，每个类别内的数据点相似性尽可能大[1]。

设有 n 个数据样本 $X = \{x_1, x_2, \cdots, x_n\} \in R^d$ 为待聚类数据集，d 维向量 $x_j = (x_{j1}, x_{j2}, \cdots, x_{jd})^{\mathrm{T}}$。寻找一个包含 k 个聚类中心的集合 $C = (c_1, c_2, \cdots, c_k)^{\mathrm{T}}$，并最小化目标函数，公式(4)：

$$J(X,C) = \sum_{i=1}^{k} \sum_{x_j \in s_i} d(x_j, c_i), \quad (4)$$

其中，S_i 是第 i 个类别中样本集，c_i 是 S_i 内所有样本 x_j 的聚类中心点，$d(x_j, c_i)$ 为样本数据 x_j 与聚类中心 c_i 之间的欧氏距离，其定义如公式(5)：

$$d(x_j, c_i) = \| x_j - c_i \|_2 = \left(\sum_{l=1}^{d} \omega_{jl} \, |x_{jl} - c_{il}|^2 \right)^{\frac{1}{2}}, \quad (5)$$

而

$$c_i = \frac{1}{n_i} \sum_{x_j \in c_j} x_j. \quad (6)$$

其中，c_i 为第 i 个类的中心位置，$i = 1, 2, \ldots, k$，x_j 代表属于类 c_i 中的样本数据，n_i 是类 c_i 中样本数据的个数。

如图 2 所示，k-means 聚类算法的流程图。

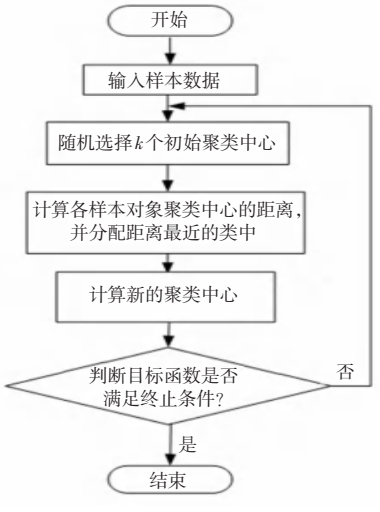

图 2　k-means 聚类算法流程图

Fig. 2　Flow chart of K-means clustering algorithm

步骤 1　每个数据样本为一个初始聚类中心，随机选取 k 个样本数据，初始聚类中心的集合为 $C = (c_1, c_2, \cdots, c_k)^{\mathrm{T}}$。

步骤 2　根据欧氏距离公式(7)，从每个数据样本到每个聚类中心的距离计算每个剩余样本数据，并划分为距离最小的类别。

$$d(x_j,c_i) = \| x_j - c_i \|_2 = \left(\sum_{l=1}^{d} \omega_{jl} |x_{jl} - c_{il}|^2 \right)^{\frac{1}{2}}, \tag{7}$$

步骤 3　公式(8)，重新进行计算 k 个聚类中心的值。

$$c_i = \frac{1}{n_i} \sum_{x_j \in c_j} x_j. \tag{8}$$

步骤 4　若目标函数公式(9)最小或保持不变，则迭代结束。

$$J(X,C) = \sum_{i=1}^{k} \sum_{x_j \in s_i} d(x_j,c_i). \tag{9}$$

3　基于差分进化的加权 k-means 算法

k-means 算法对初始聚类中心的选择很敏感，为了解决异常点使聚类结果不稳定问题，采用差分进化算法选择最佳数据点，并确定初始聚类中心。给每个样本赋予一个权重值根据每个样本的重要性，得到加权欧氏距离，增加数据属性间的区分度，来减少异常点等造成的不利影响[2]。

3.1　基于差分进化的初始聚类中心选择

随机从数据样本中选择一组数据作为初始聚类中心，构建初始种群进行编码。对差分进化算法变异、交叉和选择操作，推导得到最佳个体。通过对最佳个体进行解码，从而得到 k-means 聚类的最佳初始聚类中心[3]。

基于差分进化的初始聚类中心选择流程如图 3 所示。

算法的具体步骤：

(1)初始化群体。假设样本数据为 d 维，种群的每个个体是 $k \times d = D$ 维向量。从样本数据中随机选取 k 个样本作为一组聚类中心，进行重复执行 N_p 次，构造初始种群通过实数编码[4]。每个个体包含一组 k 个聚类中心，N_p 为种群规模，代表 N_p 个个体，在初始化种群时，取进化代数 $g = 0$，编码方式如式(10)所示：

$$X_j(0) = (x_{j1}, x_{j2}, \cdots, x_{jk}) \tag{10}$$

其中 $j = 1,2,\cdots, N_p$，$X_j(0)$ 表示初始种群的第 j 个个体，$X_{ji}(i = 1,2,\cdots,k)$ 表示第 j 个体的第 i 个聚类中心。

(2)变异操作。一个聚类中心相当于一个基因位置，变异操作是根据个体的基因位置进行的，从当前种群 $X_j(g)$ 中随机选取 3 个个体，即 $X_a(g)$，$X_b(g)$，$X_c(g)$，且 $a \neq b \neq c \neq j$，变异个体 $\nu_j(g') = (\nu_{j1}(g'), \nu_{j2}(g'), \cdots, \nu_{jD}(g'))$，且 $g' = g + 1$，种群中的每个个体的基因位如公式(11)所示：

$$\nu_{ji}(g+1) = x_{ai}(g) + \alpha(x_{bi}(g) - x_{ci}(g)). \tag{11}$$

其中，$i = 1,2,\cdots,k, a \in [0,1]$ 为缩放系数。

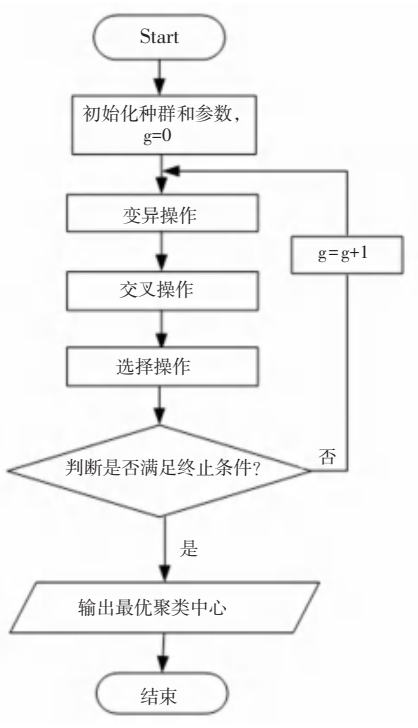

图 3　基于差异进化的初始聚类选择

Fig. 3　Initial cluster selection based on differential evolution

(3)交叉操作。当前个体 $X_j(g)$ 和变异个体 $\nu_{ji}(g+1)$ 进行交叉操作来获得中间个体 $M_j(g+1) = (m_{j1}(g+1), m_{j2}(g+1), \cdots, m_{jk}(g+1))$，中间个体的第 i 分量如下公式(12)：

$$m_{ji}(g+1) = \begin{cases} \nu_{ji}(g+1) \text{ if } \beta \leqslant C_R \text{ or } i = \gamma; \\ x_{ji}(g) \text{ else.} \end{cases} \tag{12}$$

其中 C_R 为交叉概率，且 $C_R \in [0,1]$，γ 为 $[1,k]$ 之间随机产生的一个整数，β 为 0 ~ 1 间满足均匀分布且随机产生的一个数。

(4)选择操作。采用贪婪算法选择进入下一代群体的个体，比较当前进化个体 $X_j(g)$ 与其对应的中间试验个体 $M_j(g+1)$ 的适度值，公式(13)。

$$X_j(g+1) = \begin{cases} X_j(g) \text{ if } f(X_j(g)) > f(M_j(g+1)); \\ M_j(g+1) \text{ else.} \end{cases} \tag{13}$$

(5)算法终止判断。进行检验种群 $X(g+1)$ 中的个体，若满足算法终止条件，输出最佳个体，否则返回到第二步继续操作，一直到输出最佳个体为止。

3.2　加权欧几里德距离

k-means 聚类分析中，每个样本数据对聚类结果的影响程度不同，虽然差分进化算法可以为k-means算法选择更合适的初始聚类中心，由于异常点的存在，聚类结果仍会受到很大影响，k-means聚类算法没有考虑到每个样本对聚类结果的影响[5]。为了解决这一问题，本文提出了一种加权的欧氏距离，根据每个样本的重要性给每个参与聚类的样本赋予一个权重值，从而减少异常点等因素对聚类结果的影响[6]。

定义权值 $\omega = [\omega_1, \omega_2, \cdots, \omega_n]^{\mathrm{T}} \in R^{n\times d}$，d 维向量：$\omega_j = [\omega_{j1}, \omega_{j2}, \cdots, \omega_{jd}]^{\mathrm{T}}$。对公式(7)引入权值 ω 来加以区分对每个样本数据与聚类中心之间的关系，改进后为公式(14)

$$d_\omega(x_j, c_i) = \| x_j - c_i \|_\omega = \left(\sum_{l=1}^{d} \omega_{jl} \, |x_{jl} - c_{il}|^2 \right)^{\frac{1}{2}}. \tag{14}$$

改进后的目标函数为公式(15)：

$$J(X, C) = \sum_{i=1}^{k} \sum_{x_j \in s_i} d_\omega(x_j, c_i) = \sum_{i=1}^{k} \sum_{x_j \in s_i} \left(\sum_{l=1}^{d} \omega_{jl} \, |x_{jl} - c_{il}|^2 \right)^{\frac{1}{2}}. \tag{15}$$

其中 $X = \{x_1, x_2, \cdots, x_n\} \in R^d$ 为待聚类数据集，$x_j = (x_{jl}, x_{j2}, \cdots, x_{jd})^{\mathrm{T}}$ 为 d 维向量。$C = (c_1, c_2, \cdots, c_k)^{\mathrm{T}}$ 为 k 个聚类中心的集合。

权值用公式(16)定义为：

$$\omega_{jl} = \frac{x_{jl}}{\frac{1}{n} \sum_{j=1}^{n} x_{jl}}. \tag{16}$$

x_{jl} 为第 j 个样本的第 l 个分量，$\frac{1}{n} \sum_{j=1}^{n} x_{jl}$ 为样本数据集中每个数据对象的第 l 个分量之和的平均值，ω 是一个反映样本数据整体分布特性的权值。

为了保持形式与 k-means 算法的欧氏距离公式一致，对公式(15)经过变换以获得公式(17)：

$$d_\omega(x_j, c_i) = \left(\sum_{l=1}^{d} \omega_{jl} \, |x_{jl} - c_{il}|^2 \right)^{\frac{1}{2}} = \left(\sum_{l=1}^{d} |u_{jl} - \mu_{jl}|^2 \right)^{\frac{1}{2}}. \tag{17}$$

其中，$u_{jl} = x_{jl}\sqrt{\omega_{jl}}$，$\mu_{jl} = c_{jl}\sqrt{\omega_{jl}}$.

3.3　基于差分进化的加权 k-means 算法的步骤

该算法需要确定聚类数 k，在基于差分进化选择初始聚类中心之前，选择固定参数：种群规模 $N_p \in [5D, 10D]$，缩放因子 $\alpha \in [0.5, 1]$，交叉概率 $C_p \in [0.8, 1]$。

基于差分进化的加权 k-means 算法流程如图 4 所示。

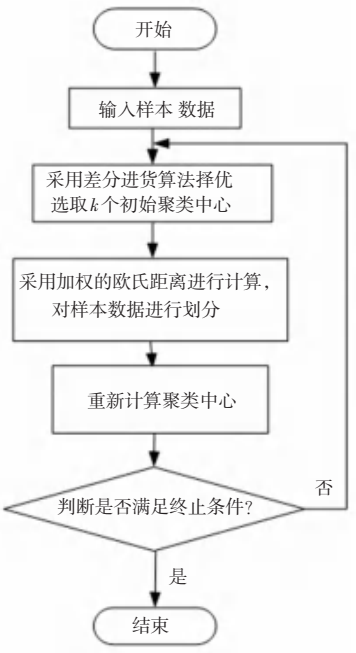

图 4　基于差分进化的加权 k-means 算法流程图

Fig. 4　Flow chart of weighted k-means algorithm based on differential evolution

算法的步骤描述：

步骤 1　输入 n 个样本的数据、聚类个数 k 和控制参数 N_p、α、C_p。

步骤 2　采用基于差分进化的初始聚类中心的选择算法来进行选取初始聚类中心。

步骤 3　根据改进的欧氏距离公式(17)，计算每个数据样本到各聚类中心的距离，将每样本划分到最小距离的类别当中。

步骤 4　根据公式(8)，重新计算 k 个聚类中心的值。

步骤 5　若满足改进的目标函数公式(15)最小或保持不变，迭代结束；否则，返回步骤 3 继续执行。

4　实验仿真与结果分析

在 UCI 实验本文选择 4 个数据集。其中，数据集名称、类别数目、样本总数和属性维数见表 1。为了说明本文提出的算法的有效性，首先利用机器学习领域的数据集对四组实验数据进行了分析和比较[7-8]。实验结果表明，样本数据集中包含的每个数据的属性维数、样本总数和类别数都相等。与 Uk-means算法、k-means 算法的聚类结果比较，算法迭代次数、聚类精度和收敛时间的比较结果见表 2~表 4。

表 1 四组 UCI 数据集

Tab. 1 Four sets of UCI data sets

数据集名称	属性维数	类别数目	样本总数
Iris	4	3	150
Wine	13	3	177
Seed	7	3	211
PageBlocks	11	5	5472

表 2 3 种方法的迭代次数

Tab. 2 Iterations of three methods

算法	迭代次数		
	k-means	Uk-means	加权 k-means
Iris	8	5	3
Wine	11	10	8
Seed	10	8	7
PageBlocks	35	22	16

表 3 3 种方法的聚类精度

Tab. 3 Clustering accuracy of three methods

算法	聚类精度/%		
	k-means	Uk-means	加权 k-means
Iris	88.67	89.21	90.05
Wine	57.32	70.06	70.21
Seed	87.63	89.04	89.01
PageBlocks	73.81	79.55	80.03

表 4 3 种方法的收敛时间

Tab. 4 Convergence time of three methods

算法	收剑时间/ms		
	k-means	Uk-means	加权 k-means
Iris	26.47	19.21	15.35
Wine	32.32	29.56	25.81
Seed	28.123	23.344	22.561
PageBlocks	469.981	315.955	221.813

在保证更好聚类精度的前提下,对于这四组不同的数据集,该算法收敛速度提高得更明显。从表 2~表 4 的实验仿真结果可以看出,该算法对四组 UCI 数据集进行聚类,能够获得较好的聚类结果。算法对目标函数的欧氏距离判别公式增加了权值,使得一些异常点与聚类中心间的欧氏距离增加,算法的每次迭代更接近真实数据的划分,分布不明显且难以分类的数据更有利于聚类,提高了聚类精度,减少了算法的迭代次数,加快了收敛速度。

5 结束语

本文提出了一种基于差分进化的加权 k-means 算法,并进行了实验仿真和结果分析。采用差分进化算法优先选择初始聚类中心,在保证聚类中心选择多样性的前提下,使初始聚类中心的选择更接近最终聚类中心;根据每个样本参与聚类的重要性,给出每个样本通过权值得到加权欧氏距离,增加了数据属性间的差异,降低了异常点对聚类结果的影响。同与其他算法比较,该算法选择的初始聚类中心更接近最终的聚类中心,保证聚类精度的同时提高了计算效率。

参考文献

[1] 王康. k_means 聚类算法的改进研究及其应用[D]. 大连理工大学,2014.
[2] 李荟娆. K_means 聚类方法的改进及其应用[D]. 东北农业大学,2014.
[3] 董明刚,王宁,程小辉. 改进的组合差分进化优化算法[J]. 计算机仿真,2013,30(1):389-392.
[4] 欧陈委. K-均值聚类算法的研究与改进[D]. 长沙理工大学,2011.
[5] 刘莉莉. K_均值聚类算法的研究与改进[D]. 曲阜师范大学,2015.
[6] 姜立强,强洪夫. 带基向量种群的改进差分进化算法[J]. 计算机工程,2012,38(3):9-11.
[7] 刘莉莉,曹宝香. 基于差分进化算法的 K-Means 算法改进[J]. 计算机技术与发展,2015(10):88-92.
[8] 乔艳霞,邹书蓉,张洪伟. 基于 K-means 的改进差分进化聚类算法[J]. 四川理工学院学报(自然科学版),2014(9):64-67.

(上接第 237 页)

同工作,可以有效地防止非机动车和行人闯红灯,从而减少交通事故的发生和人员伤亡。PLC 技术编程简便,开发周期短,可以很简单地进行代码移植以及安装,可同时实现多个交叉口信号灯与栅栏的协同作用。当机动车在交叉口发生碰撞时,栅栏可以在一定程度上保护导流岛上的行人。栅栏配合信号灯,可实现高峰配时、绿色通道等功能。如果将该项目应用于现实生活中,可以有效减少行人和非机动车闯红灯造成的交通事故,减少直接经济损失。

参考文献

[1] 杨一凡,吴崇远. 看完 2018 年杭州电动车交通事故大数据,就知道戴安全头盔有多重要[N/OL]. 钱江晚报,2019-01-23. https://baijiahao.baidu.com/s?id=1623436567233593150&wfr=spider&for=pc.
[2] 余亮. 一种行人闯红灯自动预警及显示曝光系统[J]. 有色冶金设计与研究,2019,40(3):27.
[3] 张凡,吕卉焘,沈小燕,等. 计划行为理论下外卖配送员闯红灯行为研究[J]. 中国安全科学学报,2019,29(5):1.
[4] 陈小红,张协奎. 单进口放行方式下交叉口行人嵌套相位设计[J]. 广西大学学报(自然科学版),2018,43(3):1211.
[5] 裴莹莹. 基于视频图像处理的车辆闯红灯违章的检测方法研究[D]. 长春:吉林大学,2018.
[6] 韩宝睿,丁莉莎,李颖,等. 前车遮挡造成的后车误闯红灯机理及信号灯设置研究[J]. 重庆交通大学学报(自然科学版),2020,39(1):23.
[7] 刘岩,付川云. 机动车闯红灯行为类型划分及特征分析[J]. 昆明理工大学学报(自然科学版),2018,43(4):115.
[8] 杨金龙. 桥式起重机控制系统研究与应用[D]. 天津:天津大学,2017.