DOI: 10.7641/CTA.2017.60645

高炉料面的分类与案例匹配算法

曹　铭, 张　森†, 尹怡欣, 肖文栋

(北京科技大学 自动化学院, 北京 100083)

摘要: 本文基于改进的k-means算法及分级案例库匹配技术提出一种研究高炉料面和煤气流关系的方法. 为了获取煤气流分布情况, 首先提出了改进的基于新型有效性指标评价的k-means算法, 并将该算法与其它多种算法进行比较, 证明了该方法的高效性和准确性. 继而在此基础上提出了案例库匹配技术, 从而获得与当前料面最为匹配的历史料面. 最后, 将匹配算法与改进的灰色相似性匹配算法和欧式近邻匹配算法进行比较. 结果表明, 分级匹配算法具有更高的分辨率和效率. 在多次试验中匹配准确率高达92.5%, 比其他几种算法更加准确, 更适合研究料面与煤气流关系, 指导布料操作.

关键词: 高炉; 料面; 煤气流; 聚类算法; 案例匹配

中图分类号: TP181　　**文献标识码**: A

Classification and case matching algorithm on blast furnace burden surface

CAO Ming, ZHANG Sen†, YIN Yi-xin, XIAO Wen-dong

(School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: A method is supposed in this paper based on the improved k-means algorithm and graded case based matching method to study the relationship between the burden surface and the gas flow in blast furnace. To get the gas flow distribution from historical data, first of all, an improved k-means algorithm was proposed based on a new evaluation approach of effectiveness index. Comparison with other clustering algorithms proved that the proposed algorithm has a high accuracy and high efficiency. A matching technique was put forward on the basis of the above clustering results to obtain the most matched historical burden surface. At last, compared with the improved grey similarity matching algorithm and Euclidean nearest neighbor matching algorithm, the results showed that the proposed method has higher resolution and efficiency. Matching accuracy is as high as 92.5% in the experiments which is more accurate than the other methods. The approach is more suitable for the investigation of the relationship between burden surface and gas flow so as to assist monitors of blast furnace to control burden surface.

Key words: blast furnaces; burden surface; gas flow; clustering algorithms; case matching

1　引言(Introduction)

钢铁是现代社会必不可少的原材料. 我国对钢铁的需求量十分巨大. 资料显示, 多年来我国是最大的钢铁生产国与消费国[1] . 但是技术的不足严重影响钢铁的产量和质量. 因此, 提高技术是急需解决的问题之一.

高炉生产中, 将铁矿石、焦炭及其它辅助原料自炉顶装入高炉, 并向炉内鼓入热风. 焦炭与氧气反应生成一氧化碳和氢气. 原料、燃料随着炉内熔炼等过程的进行而下降, 炉料和煤气相遇, 发生传热、还原、熔化、脱炭生成生铁. 这个过程中, 炉料与煤气流相互作用相互影响. 煤气流分布反映炉况顺行与否, 料面是易测量的一个变量. 通过料面得到对应的煤气流分布, 对于提高我国高炉炼铁水平尤为重要.

煤气流分布主要与料面有关. 但是, 获取料面和煤气流关系的研究特别少, 这也成为本文研究的一大亮点与创新点. 本文基于聚类匹配算法研究高炉料面和煤气流关系, 进而辅助布料. 思路如图1所示. 首先, 建立历史案例库, 保存料面数据. 选取足够长时间内测得的数据作为历史案例库. 然后, 利用本文提出的类内类间可调节比率(inner-class inter-class adjustable ratio, ICICAR)算法, 对料面数据进行聚类, 得到聚类中心即代表案例和不同类别的子案例库. 接着, 利用本文的RM (ranking matching)匹配算法将当前料面与代表案例匹配, 再与对应的子代案例匹配, 获得最匹配的历史料面. 而历史料面对应的煤气流分布已知,

收稿日期: 2016−8−29; 录用日期: 2016−12−28.

†通信作者. E-mail: zhangsen@ustb.edu.cn; Tel.: +86 15810580275.

本文责任编委: 李少远.

国家自然科学基金重点项目(61673056, 61333002, 61673055)资助.

Supported by National Natural Science Foundation of China (61673056, 61333002, 61673055).

通过查阅历史纪录, 得到对应的历史煤气流分布, 即当前煤气流分布. 先与代表案例匹配再与子代案例匹配, 可以有效提高效率. 高炉历史料面数据规模极其庞大, 假设有一百万条. 如果直接匹配就需要匹配一百万次, 而如果按本文思路, 假设聚为十类, 每类十万条, 那么先进行代表案例匹配再与子代案例匹配则只需要匹配十万零一次. 本方法跳过函数关系的求解, 使问题简单化.

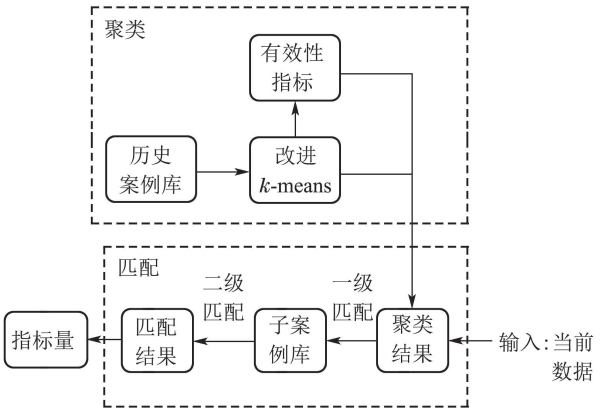

图 1 案例匹配示意图

Fig. 1 Case matching diagram

2 ICICAR聚类算法(ICICAR clustering algorithm)

2.1 改进的 k-means 算法 (The improved k-means algorithm)

传统k-means[2]算法逻辑简明清晰, 效率高, 被普遍应用. 但是, 传统的k-means算法必须给定聚类数. 这是k-means的一个缺点. 再者, k-means依赖初始聚类中心, 而初始聚类中心往往是随机选定的, 这就使得聚类结果具有不确定性. 而且k-means将该类中所有对象的均值作为聚类中心, 结果受到孤立点的很大影响. 针对上述缺点, 本文在传统的k-means算法上进行了一些改进.

对于初始聚类中心的选取, 首先选出所有样本中距离最大的两个点, 然后选出与已选出的点距离最大的另一个点, 直至选出k个点, 作为初始聚类中心. 对于孤立点, 在迭代过程中选取第k代聚类中心时, 将与第$k-1$代聚类中心相似度较大的该类样本作为子集计算均值, 作为第k代聚类中心. 这样可以将孤立点作为聚类中心的情况排除在外, 使聚类中心向簇中密集区靠拢, 减小孤立点的影响.

2.2 有效性指标(Effectiveness index)

聚类的有效性指标是指选取的评价聚类好坏的标准. 本文提出一种新的有效性指标, 用于确定聚类数. 首先假设数据集可以分成k类, 具体表示为$C^k=\{C_1, C_2, \cdots, C_k\}$. 本文用$\mathrm{Inner}(C^k)$表示类内紧凑性, 用$\mathrm{Inner}(C^k)$表示类间分离度. 具体定义如下:

$$\mathrm{Inner}(C^k)=\frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mathrm{centr}(i)\|, \tag{1}$$

$$\mathrm{Inner}(C^k)=\frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\|\mathrm{centr}(i)-\mathrm{centr}(j)\|, \tag{2}$$

其中: $\mathrm{centr}(i), \mathrm{centr}(j)$表示$k$类中第$i$, j类数据的聚类中心, n表示样本数. $\mathrm{Inner}(C^k)$越小, 说明类内样本点越紧凑; $\mathrm{Inter}(C^k)$越大, 说明各类之间越分离.

本文用如下表达式确定聚类效果:

$$Q(C)=\mathrm{Inner}(C^k)-\beta\cdot\mathrm{Inter}(C^k), \tag{3}$$

β反映对类内紧凑性和类间分离度的重视程度$Q(C)$越小, 聚类效果越好.

而过去常用的一种有效性指标方法是令

$$\mathrm{Inner}(C^k)=\sum_{i=1}^{k}\sum_{X,Y\in C_i}\|X-Y\|^2, \tag{4}$$

X, Y是类C_i中的任意两个数据对象.

$$\mathrm{Inner}(C^k)=\sum_{i=1}^{k}\Big(\sum_{j\neq i,j=1}^{k}\frac{1}{|C_i|\cdot|C_j|}\sum_{X\in C_i,Y\in C_j}\|X-Y\|^2\Big), \tag{5}$$

其中: X, Y分别属于类C_i, C_j中的两个数据对象$|C_i|$, $|C_j|$表示类C_i, C_j中数据对象的个数.

本文实验数据为某钢厂提供的料面数据炉顶装有一个角度可调的雷达测量十次得到十个点的料面数据, 然后, 对十点雷达数据进行插值获得31点数据, 以此为研究对象. 各维是以炉心为原点, 横坐标为-4.05, -3.56, -3.08, -2.59, -2.43, -2.26, -2.09, -1.88, -1.70, -1.49, -1.24, -0.96, -0.74, -0.49, -0.25, 0, 0.25, 0.49, 0.74, 0.96, 1.24, 1.49, 1.70, 1.88, 2.09, 2.26, 2.43, 2.59, 3.08, 3.56, 4.05, 对应料面到零料线的距离. 历史案例库容量为3225, 该钢厂专家一致认为分为6类, 历史案例库的数据结构可以表示为

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,30} & x_{1,31} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,30} & x_{2,31} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,30} & x_{i,31} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{3225,1} & x_{3225,2} & \cdots & x_{3225,30} & x_{3225,31} \end{bmatrix},$$

其中$x_{i,j}$表示第i个数据的第j个属性值. 本文聚类数确定仿真见图2所示.

而对另一种方法进行仿真, 仿真结果见图3所示.

图2–3中, 横坐标表示不同的聚类数, 纵坐标表示对应的评价指标. 评价指标值越小, 聚类越优. 从图中可以看到本文算法确定的聚类数为6, 一般方法的聚类数为50, 这也验证了本文提出的聚类数确定方法对

于高炉料面聚类数的判断上的可行性, 以及较之一般方法有一定优势.

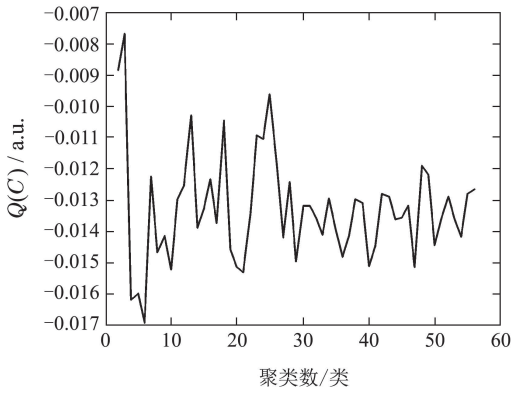

图 2 高炉料面聚类数确定

Fig. 2 The clustering number of burden surface

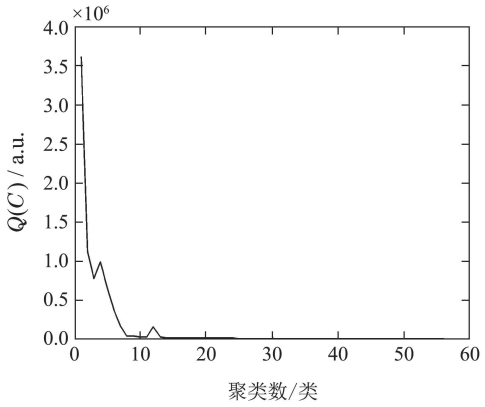

图 3 普通方法聚类数确定

Fig. 3 The clustering number of another method

2.3 ICICAR算法实现流程(Implementation process of ICICAR algorithm)

ICICAR算法要求反复迭代改进的k-means算法, 并分别求出各自的有效性指标, 据此确定最佳聚类数.

在文献[3]中给出了聚类数范围, $k_{\min}=1$, $k_{\max}=\text{int}(\sqrt{n})$, 其中$n$为样本数. 对前面提到的31维高炉料面数据进行聚类, 聚类结果如图4所示.

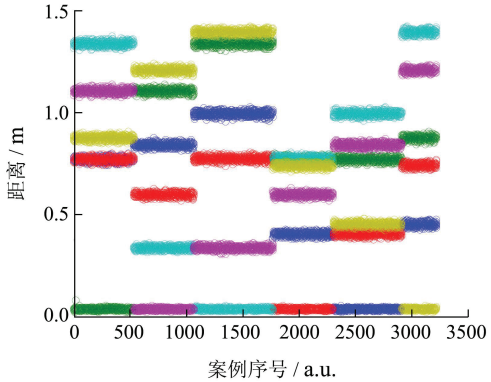

图 4 高炉料面聚类结果

Fig. 4 The clustering result of burden surface

图4中, 横坐标是历史案例库中具体的案例索引, 纵坐标是历史料面到料面聚类中心的距离. 不同的颜色对应不同的类. 由此可得出各类的聚类半径, 如图5所示.

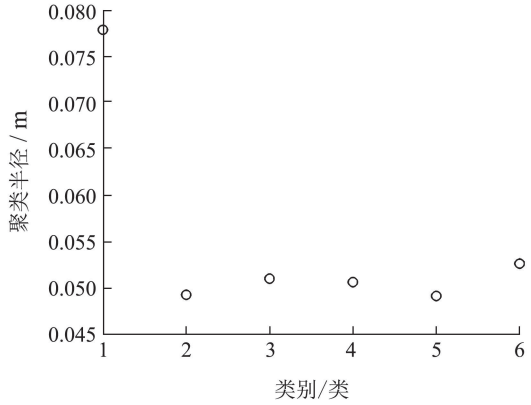

图 5 高炉料面聚类半径

Fig. 5 The clustering radius of burden surface

2.4 ICICAR算法与其他算法的比较(Comparison of ICICAR algorithm with others)

高炉料面形状一般为“草帽型”, 与正态分布形状类似. 料面数据量庞大, 雷达测量方式快速方便, 为了准确确定料面形状以及煤气流的分布关系, 算法的效率尤为重要. 本节以高炉料面为背景, 将ICICAR与其它几种算法进行比较, 进行100次实验, 结果见表1.

表 1 聚类方法比较

Table 1 Comparison of clustering algorithms

聚类算法	准确率/%	耗时/s
FCM	83.63	4.0652
单连通	78.54	1.6094
谱聚类	85.46	3.8374
广度优先邻居搜索	83.38	1.5625
ICICAR算法	85.77	0.9386

可以看出, ICICAR算法的准确率略高于其它几种, 它的效率远高于其他算法. 高炉现场的数据量非常庞大, 这样总的计算耗时就会成几何倍增加. 这样, ICICAR算法效率就更为突出, 更适合于高炉实际生产.

3 RM匹配算法(RM matching algorithm)

对于案例库匹配, 前人做过许多研究. Li-min Xia[4]等为了度量时间顺序的相似性, 提出一种基于顺序因子的相似性度量方法. EsmatRashedi等[5]在基于内容的图像检索(content based image retrieval, CBIR)系统中采用案例推理的长期学习方法. Weng Min等[6]采取案例推理的方式进行路径规划. Zhou P等[7]利用基于案例推理和模糊相似度粗糙集的数据驱动软传感器

进行产品质量估计. Eduardo Lupiani等[8]提出一种用于评价案例库维护算法的具体方法. Zheng LU等[9]提出结合社区发现和案例推理的混合方式, 解决当前工业条件下测量困难的问题. Ying Lu等[10]利用案例推理技术分析安全风险问题. Rosanne Janssen等[11]用案例推理方法预测焦虑症患者的治疗效果. Morteza-Behbahani等[12]给出了统计过程表示案例的一种格式和案例检索的相似性度量. 但是, 高炉生产中很少应用匹配算法. 下面对本文提出的RM匹配算法进行说明.

3.1 相似度计算(Similarity calculation)

为了使相似度具有更好的精度和区分能力, 采用下述定义. 设案例$x_i, x_j \in U(1 \leqslant i, j \leqslant n)$, n表示案例规模, 在属性$q_k \in A$上, x_i, x_j取值为a_{ik}, a_{jk}, 属性q_k在U上的最大值和最小值分别为$b_{k\max}, b_{k\min}$, ϑ为属性相似度阈值, 那么, 案例x_i, x_j在属性q_k上的相似度为

$$\mathrm{sim}_{q_k}(x_i, x_j) = \begin{cases} \mathrm{sim}'_{q_k}(x_i, x_j), & \text{如果 } \mathrm{sim}'_{q_k}(x_i, x_j) \geqslant \vartheta, \\ 0, & \text{其他}, \end{cases} \tag{6}$$

则总体相似度为

$$\mathrm{sim}_A(x_i, x_j) = \sum_{k=1}^{m} w_k \mathrm{sim}_{q_k}(x_i, x_j) / \sum_{k=1}^{m} w_k, \tag{7}$$

w_k为第k个属性对应的权值. 料面数据属性相似度, 即料面单维数据相似度, 在一定程度上降低了高炉复杂环境对于雷达测量的干扰. 整体相似度反映当前数据与历史案例的相似程度.

3.2 权值的确定(The determination of weights)

首先引入覆盖度的定义. 设案例$x_i \in U$, x_i的覆盖度是指案例库U中与x_i相似的所有案例的集合, 即

$$\mathrm{cover}(x_i) = \{x_j | x_j \in U,\ \mathrm{sim}_A(x_i, x_j) \geqslant \alpha\}, \tag{8}$$

式中: A为属性集, 这里指料面案例库数据指定维的集合, α为案例相似度阈值, $\mathrm{sim}_A(x_i, x_j)$表示x_i, x_j关于A的相似度. 这里权值确定基于覆盖度. 假设权值相等, 比较不同属性对平均覆盖度的影响, 从而计算每个属性的权值.具体步骤如下表示:

Step 1　对案例进行简单排列, 然后令权值为$W_k = 1/m (k = 1, 2, \cdots, m), i = 1, m$表示属性个数;

Step 2　计算含所有属性时x_i的覆盖度, 覆盖案例个数记为C_0, 并令$k = 1$;

Step 3　计算去掉属性q_k时x_i覆盖度, 覆盖案例个数记为C_{ki};

Step 4　判断$k \geqslant m$, 若成立, 则继续执行; 否则, 令$k = k + 1$, 跳到Step 3;

Step 5　判断$i \geqslant n$, 若成立, 则继续执行; 否则, 令$i = i + 1$, 跳到Step 2;

Step 6　根据不同属性对案例平均覆盖度的影响程度, 确定特征属性q_k的重要性w'_k:

$$w'_k = \frac{|\dfrac{1}{n}\sum_{i=1}^{n} C_{0i} - \dfrac{1}{n}\sum_{i=1}^{n} C_{ki}|}{\dfrac{1}{n}\sum_{i=1}^{n} C_{0i}},\ k = 1, 2, \cdots, m; \tag{9}$$

Step 7　对w'_k归一化, 则可得到每个特征属性的权值w_k:

$$w_k = \frac{w'_k}{\sum_{k=1}^{m} w'_k},\ k = 1, 2, \cdots, m. \tag{10}$$

实际生产中, 不同高炉的权值不同. 而且根据当前的经验, 工作人员也很难给出一个主观参考. 因此, 本文直接对客观数据进行研究确定权值.

3.3 RM算法与其他算法的比较(Comparison of RM algorithm and other algorithms)

首先引入相似度分辨率的概念. 相似度分辨率, 即样本数据变化较小时, 输入数据与各样本相似度差异的大小. 差异越大, 分辨率越高; 差异越小, 分辨率越低. 将当前料面数据与代表案例库进行匹配, 匹配结果见表2(K表示代表案例).

表 2 匹配算法分辨率和速度的比较

Table 2 Comparison of resolution and speed of matching algorithm

匹配算法	当前数据与代表案例的相似度						平均所需时间/s
	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	
灰度方法	0.9215	0.3425	0.3278	0.4250	0.3737	0.2648	1.3438
欧式近邻方法	0.9682	0.2768	0.0516	0.3007	0.3269	0.3371	0.0601
RM方法	0.9743	0.1611	0	0.5044	0.3858	0	0.0625

已知匹配料面属于第1类. 可以看出, 改进的灰度匹配算法耗时是其它算法的20多倍, 而且分辨率很低. 如果代表案例较多的时候, 时间成本高, 不适用于高炉生产. 而欧式近邻算法与RM算法相比, 所需时间接近, 但是对于较接近的案例, 欧式近邻方法分辨率较低, 而RM方法区分程度较高. 当与料面的子案例库匹

配时, 子代案例个数平均有500多个. 改进的灰度算法所需时间太长, 而欧式近邻方法需要99 s, 而RM方法需要120 s. 效率上, 欧式近邻方法最高, 其次RM方法, 改进灰度方法效率最低. 然后, 比较匹配准确性. 从钢厂公认的6类料面数据中每类任意取出20个料面与整个历史案例库进行匹配测试, 匹配结果见表3.

表 3 匹配准确性比较

Table 3 Comparison of the accuracy of matching algorithm

匹配算法	测试次数	正确次数
改进灰度方法	120	88
欧式近邻方法	120	84
RM算法	120	111

可以看出, RM算法准确率更高, 所以综合考虑采用RM匹配算法较为合适. 获得匹配结果后, 查阅历史纪录, 匹配到的料面案例对应的煤气流分布即为当下煤气流分布.

4 结论(Conclusions)

本文提出一种研究高炉料面与煤气流关系的方法. 创新点主要有两方面: 一是研究问题的创新. 查阅国内外关于料面与煤气流分布的研究, 发现对二者关系的研究非常少, 本文为以后的研究提供思路. 二是将案例库思想、聚类、匹配方法应用到料面与煤气流分布的研究中, 对算法进行改进, 结果与专家判断基本相符, 具有很强的实用价值.

本文提出基于改进k-means和新型有效性指标的ICICAR聚类算法, 与其他算法相比, 该算法在高炉生产中就聚类数确定、聚类效率和准确性而言表现更好. 然后提出RM匹配算法, 该算法可以高效准确的获得匹配到的历史料面. 并且通过实验比较了该算法与其它算法的性能, 证明RM算法的高效性、准确性和可行性.

参考文献(References):

[1] LIU Dexin, LI Xiaoli, DING Dawei, et al. Multi-model control of blast furnace burden surface based on observed data of radars [J]. *Control Theory & Applications*, 2012, 29(10): 1277 – 1283.
(刘德馨, 李晓理, 丁大伟, 等. 基于雷达观测数据的高炉料面多模型控制 [J]. 控制理论与应用, 2012, 29(10): 1277 – 1283.)

[2] LI H Y, HE H Z, WEN Y G. Dynamic particle swarm optimization and k-means clustering algorithm for image segmentation [J]. *Optik-International Journal for Light and Electron Optics*, 2015, 126(24): 4817 – 4822.

[3] CHEN Lifei, JIANG Qingshan, WANG Shengrui. Ahierarchical method for determining the number of clusters [J]. *Journal of Software*, 2008, 19(1): 62 – 72.
(陈黎飞, 姜青山, 王声瑞. 基于层次划分的最佳聚类数确定方法 [J]. 软件学报, 2008, 19(1): 62 – 72.)

[4] XIA L M, YANG B J, TU H B. Recognition of suspicious behavior using case-based reasoning [J]. *Journal of Central South University*, 2015, 22(1): 241 – 250.

[5] ESMAT R, HOSSEIN N P, SAEID S. Long term learning in image retrieval systems using case based reasoning [J]. *Engineering Applications of Artificial Intelligence*, 2014, 35(10): 26 – 37.

[6] WENG M, WEI X Q, QU R, et al. A path planning algorithm based on typical case reasoning [J]. *Geo-spatial Information Science*, 2009, 12(1): 66 – 71.

[7] ZHOU P，LU S W，CHAI T. Data-driven soft-sensor modeling for product quality estimation using case-based reasoning and fuzzy-similarity rough sets [J]. *IEEE Transactions on Automation Science & Engineering*, 2014, 11(4): 992 – 1003.

[8] EDUARDO L, JOSE M J, JOSE P. Evaluating case–base maintenance algorithms [J]. *Knowledge-Based Systems*, 2014, 67(3): 180 – 194.

[9] LÜ Z, LIU Y, ZHAO J, et al. Soft computing for overflow particle size in grinding process based on hybrid case based reasoning [J]. *Applied Soft Computing*, 2015, 27(C): 533 – 542.

[10] LU Y, LI Q M, XIAO W J. Case–based reasoning for automated safety risk analysis on subway operation: case representation and retrieval [J]. *Safety Science*, 2013, 57(8): 75 – 81.

[11] JANSSEN R, SPRONCK P, ARNTZ A. Case–based reasoning for predicting the success of therapy [J]. *Expert Systems*, 2015, 32(2): 165 – 177.

[12] BEHBAHANI M, SAGHAEE A, NOOROSSANA R. A case–based reasoning system development for statistical process control:case representation and retrieval [J]. *Computers and Industrial Engineering*, 2012, 63(4): 1107 – 1117.

作者简介:

曹 铭 (1991–), 男, 硕士研究生, 目前研究方向为高炉布料系统建模与控制, Email: cm_ustb@163.com;

张 森 (1971–), 女, 副教授, 硕士生导师, 目前研究方向为高炉布料系统建模与控制, Email: zhangsen@ustb.edu.cn;

尹怡欣 (1957–), 男, 教授, 博士生导师, 目前研究方向为高炉布料系统建模与控制, Email: yyx@ies.ustb.edu.cn;

肖文栋 (1968–), 男, 教授, 博士生导师, 目前研究方向为模式识别算法及无线传感器网络, Email: wdxiao@ustb.edu.cn.