

热轧工艺关键变量的提取方法^①

钟 蜜, 刘 斌, 蒋 峥

(武汉科技大学 信息科学与工程学院, 武汉 430081)

摘 要: 在热轧带钢生产中, 终轧温度的控制精度是影响产品质量的重要指标, 而影响精轧终轧温度的因素较多且复杂, 从单一影响因素上分析对精轧温度的影响是不够准确和全面的。因此, 提出了一种基于属性重要性的热轧关键变量提取方法。该方法是从整个决策系统的角度上分析每个可能对精轧目标温度有影响的属性, 筛选掉对目标温度属性没有影响的属性, 确定每一个属性的属性重要性大小, 再根据温度预报精度要求以及热轧工艺特点, 提取影响精轧终轧温度的关键特征变量。实际应用结果表明, 这种方法可以有效的提取出对热轧终轧温度有影响的关键特征变量。

关键词: 关键变量提取; 属性重要性; 模糊聚类; 热轧工艺

Key Variable Extraction in Hot Rolling Process

ZHONG Mi, LIU Bin, JIANG Zheng

(College of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China)

Abstract: In the process of hot strip production, the control accuracy of the finish rolling temperature is an important indicator which affects quality of the products. As factors which have impact on the final finish rolling temperature are numerous and complicated, it's inaccurate and incomprehensive to analyze by looking at one single factor. With this in mind, a key variable extraction method in the hot rolling process based on importance of attributes is presented. This method, in the perspective of the entire decision-making system, analyzes each affected attributes, filter out the temperature attribute which doesn't affect the finish rolling temperature, determines the importance of each attribute. Based on the requirement of the temperature prediction accuracy and the characteristics of hot rolling process, the system extracts the key features of variables affecting the final finish rolling temperature. The application example demonstrates that this method can effectively extract the key features of variables that affect the final finish rolling temperature.

Key words: extraction of key variables; attribute importance; fuzzy clustering; hot rolling process

在热轧带钢生产中, 精轧机组终轧温度的控制精度将直接影响到最终产品的组织性能。终轧温度精度控制对各精轧机架的轧制力和负荷分配有很重要的影响, 精轧带钢全长终轧温度控制一直是热轧生产中的重要研究课题, 也是难点之一。因此, 需建立可靠的终轧温度控制模型来对精轧终轧温度进行有效控制, 其关键前提, 就是从众多的精轧过程变量中提取出对精轧终轧温度有决定性影响的若干关键特征变量。

粗糙集理论是 Z.Pawlak 于 1982 年在波兰提出的一种智能决策数学分析工具, 广泛应用于不确定性和不完全性的信息系统进行分析和知识发现^[1]。属性重要性概念最初是在考查知识约减时被提出来的, 是粗糙集理论中的重要部分。它主要被用于抽取等式关系最小独立部分, 来对决策表进行约简, 提取特征属性。为了找出某些属性或是属性集在整个样本集中的重要性, 需要从中去掉某些属性, 看这些属性的删除是否

① 基金项目:湖北省优秀中青年创新团队项目(T200801);湖北省教育厅重点项目(D20091103);武汉科技大学教研项目(2008033)

收稿时间:2011-03-29;收到修改稿时间:2011-05-01

对样本的分类有影响,若删掉后对分类情况改变较大,则其属性重要性高^[2,3]。本文就是从整个决策系统的角度上来分析并提取出热轧工艺中影响目标温度的主要因素。实际应用表明,该方法是切实可行的且应用效果显著。

1 连续属性离散化

连续属性的离散化是数据挖掘和机器学习中比较重要的预处理步骤,在规则提取和特征分类等许多算法中,只能处理离散化的数据。常用的离散化方法主要有等区间宽度的方法和等频方法、基于熵的离散化方法以及基于聚类的离散化方法,离散化的方法不同,结果也往往存在差异^[4,5]。

由于模糊聚类算法对于处理不确定性和不完全性的信息系统有很好的分类效果^[7],因此我们选择基于模糊聚类算法来对属性进行离散化处理。

模糊聚类算法的基本原理是:首先随机选择若干个聚类中心,所有数据对象被赋予对聚类中的隶属度,然后通过迭代方法不断地修正聚类中心,迭代过程中以极小化所有的数据对象到各个聚类中心的距离与隶属度的加权和为优化目标。

模糊聚类的输入为:连续属性集 $C = \{C_1, C_2, \dots, C_i, \dots, C_s\}$ 以及每一个属性下的属性值 $C_i = \{x_1, x_2, \dots, x_j, \dots, x_n\}$;

模糊聚类的输出为:每一个属性 C_i 聚类后的中心点和每个属性值对应的离散区间符号。

算法具体步骤为:

Step1: 分别取出每一个连续属性 C_i 的属性值 $\{x_1, x_2, \dots, x_j, \dots, x_n\}$, 并进行归一化处理;

Step2: 初始化聚类中心 $c_1, c_2, \dots, c_i, \dots, c_k$, 为聚类中心个数;

Step3: 根据聚类中心计算样本之间的距离 $d_{ij} = d(x_j - c_i) = |x_j - c_i|$, 再根据该距离计算隶属度矩阵

$$u_{ij} = \frac{1}{\sum_{p=1}^k \left(\frac{d_{ij}}{d_{pj}}\right)^{2/(m-1)}}, \text{ 为加权指数, 一般取值 } 2;$$

Step4: 根据隶属度矩阵求得聚类中心 $c_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m}$;

Step5: 当迭代次数达到 20 次或者聚类目标函数评估值 $J(U, c_1, \dots, c_k) = \sum_{i=1}^k J_i = \sum_{i=1}^k \sum_{j=1}^n u_{ij}^m d_{ij}^2$ 小于阈值 ε (取

值为 0.00001), 则算法停止直接到 Step6, 否则返回到 Step3;

Step6: 根据隶属度矩阵确定每个属性的属性值对应的离散区间, 分别用 0,1,2,...,k-1 表示。

2 基于属性重要性的关键变量提取

从数据采集系统中采集的热轧过程数据, 属性条目众多, 且在整个知识库中并不是同等重要, 属性间还有可能存在冗余, 不利于做出正确而简洁的决策。因此, 在做出决策之前就需要对所有的属性进行一个筛选, 删掉对决策影响不大的属性, 挑选出对决策有重要影响的属性。由于影响带钢精轧过程中温度变化的因素很多, 且各属性间的耦合情况比较严重, 分别从某一个因素上考虑其对终轧温度的影响, 不够精确和全面, 在此基础上本文提出了基于属性重要性的热轧关键变量提取方法。

2.1 基本概念

设 $S = (U, A, V, F)$ 为一个信息系统, 其中, $U = \{x_1, x_2, \dots, x_n\}$ 是论域; A 是属性集合; V 是属性取值集合; F 是 $U \times A \rightarrow V$ 的映射。若 $A = C \cup D$, $C \cap D = \emptyset$, C 称为条件属性集, 称为决策属性集, D 则该信息系统称为决策表^[5-7]。

定义 1: 设 $S = (U, A, V, F)$ 是一个决策表, 对于每一个子集 $B \subseteq A$, 我们定义 $IND(B)$ 为不可分辨关系:

$$IND(B) = \{(x, y) | (x, y) \in U^2, \forall b \in B(b(x) = b(y))\}$$

定义 2: 设 U 为一个论域, P 和 Q 为 U 上的 2 个等价关系簇, Q 的 P 正域记为 $POS_P(Q)$, 定义为 $POS_P(Q) = \bigcup_{X \in U/Q} P_{-}(X)$ 。

定义 3: 设 U 为一个论域, $P \subseteq C$, 对于 D 的 k 个划分 $\{Y_1, Y_2, \dots, Y_k\}$, P 的近似精度为 $\gamma_P = \sum_{i=1}^k \text{card}(P_{-}Y_i) / \text{card}(U)$, 其中 $\text{card}()$ 表示集合的基数; γ_P 反映决策表分类的正确程度, 描述了关于论域的知识完备程度。

定义 4: 设 $S = (U, A, V, F)$ 是一个决策表, 条件属性子集 C , 任意条件属性 $c \in C$ 相对于条件属性集合 C 对决策属性集合 D 依赖程度的属性重要性定义为 $\text{sgf}_{C-\{c\}}^D = \gamma_C - \gamma_{C-\{c\}}$ 。

2.2 基于属性重要性的属性约减算法

对于决策信息系统 $S = (U, A, V, F)$, 其中 $A = C \cup D$, $C \cap D = \emptyset$, C 为条件属性集, D 为决策属

性集, 设 $c \in C$, 若 $sgf_{C-\{c\}}^D(c) = 0$, 则称 c 在 C 中对于 D 是不重要的, 若 $sgf_{C-\{c\}}^D(c) = 1$, 则称 c 在 C 中对于 D 是重要的, 若有 $0 < sgf_{C-\{c\}}^D(c) < 1$, 则称 c 在 C 中对于 D 是部分重要的。文中以属性重要性求取约减算法是以所有条件属性为初始约减集合, 以属性重要性 $0 < sgf_{C-\{c\}}^D(c) < \varepsilon$ 为约减标准, 逐步缩减来求取约减。其算法步骤为:

Step1: 初始化条件属性集 U/C 和决策属性集 U/D ;

Step2: 对于每一个属性 $c \in C$, 求属性 c 的属性重要性 $sgf_{C-\{c\}}^D(c) = \gamma_C - \gamma_{C-\{c\}}$;

Step3: 将属性重要性小于 ε 的属性从属性列表中删除;

Step4: 对于决策表中剩余的条件属性, 可根据其属性重要性的大小对属性进行排序。

该算法就是从条件属性集中删掉重要性比较小的属性, 再根据属性重要性大小, 结合实际轧制工艺特点, 选择出合适的条件属性集。

2.3 算法示例

表 1 原始热轧数据表

U	条件属性						决策属性
	HA RD NE SS	SPEED UP	FM_ WIDT H	RM_ TEM P	...	FM_ THIC K	D
1	4	7.894	1875.14	980	...	11.97	NW
2	4	7.41	1832.55	1000	...	11.83	best
3	6	8.99	1645.74	980	...	9.008	PW
4	5	2.868	1413.47	990	...	4.115	NB
5	5	6.076	1748.49	990	...	5.989	NB
6	7	7.98	1058.57	980	...	7.836	PB
7	3	3.384	1196.79	1000	...	3.617	PW
8	4	8.537	1388.13	1000	...	4.063	best

轧制过程数据经过预处理后, 我们选择其中的 460 条作为样本数据, 样本数据内有 45 个条件属性和一个目标属性组成。表 1 所示为部分数据, 其中离散属性 HARDNESS(带钢硬度), 其离散值有 {0,1,2,3,4,5,6,7}; 连续属性有 RM_TEMP(粗轧出口温度), FM_WIDTHH(精轧宽度), SPEED_UP(加速度), FM_THICK(精轧厚度)等。对于目标属性精轧终轧温度的离散化是以精轧终轧目标温度为中心的对称分布为五个区域: NW, NB, PB, PW, best。

原始热轧数据经过离散化后得到如表 2 所示的决策表, 由于样例集的条数比较多, 我们选择前 8 条数据作为一个小的决策表来对算法进行说明。

表 2 简单的决策表

U	条件属性				决策属性
	c1	c2	c3	c4	D
	SPEED_ UP	FM_WID TH	RM_TEM P	FM_THI CK	
1	1	0	0	2	-2
2	1	0	2	2	0
3	1	0	0	2	1
4	0	1	1	0	-1
5	1	0	1	1	-1
6	1	2	0	1	2
7	0	2	2	0	1
8	1	1	2	1	0

依据表 2 所示的决策表来计算每一个属性的重要性:

$$1) U = \{1, 2, 3, 4, 5, 6, 7, 8\}; C = \{c_1, c_2, c_3, c_4\}$$

$$U/D = \{(1), (2, 8), (3, 7), (4, 5), 6\};$$

$$U/C = \{(1, 2), (3), (4), (5), (6), (7), (8)\}.$$

其中, $Pos_C(D) = \{3, 4, 5, 6, 7, 8\}$

$$\text{则 } \gamma_C(D) = \frac{Card(Pos_C(D))}{Card(U)} = \frac{6}{8} = 0.75.$$

2) 对于属性 c_1 , 假设将属性 c_1 从条件属性集中删除, 则 $U/IND(C - \{c_1\}) = \{(1, 3), (2), (4), (5), (6), (7), (8)\}$

$$Pos_{C-\{c_1\}}(D) = \{2, 4, 5, 6, 7, 8\}; \text{ 则 } \gamma_{C-\{c_1\}}(D) = \frac{6}{8} = 0.75;$$

3) 属性 c_1 的重要性为:

$$sgf_{C-\{c_1\}}^D = \gamma_C(D) - \gamma_{C-\{c_1\}}(D) = 0;$$

设 $\varepsilon = 0.137$, 则 $sgf_{C-\{c_1\}}^D < \varepsilon$, 所以可以将 c_1 从决策表中删掉, 剩余的条件属性即为 c_2, c_3, c_4 ;

4) 继续按照步骤 2 计算条件属性 c_2, c_3, c_4 的重要性;

5) 最终输出约减后的条件数据集。

3 关键变量提取结果分析

我们选择 45 个可能对终轧温度有影响的属性作为分析对象。经过算法处理后, 得到每个属性的属性重要性大小, 即属性对终轧目标温度的影响程度, 再

将属性按照重要性大小进行排序,结果如图1所示,其中对目标温度影响大的属性表现的比较突出的,这些属性确实与目标变量终轧温度之间的有很密切的联系,这也直接反应了精轧的工艺本质。

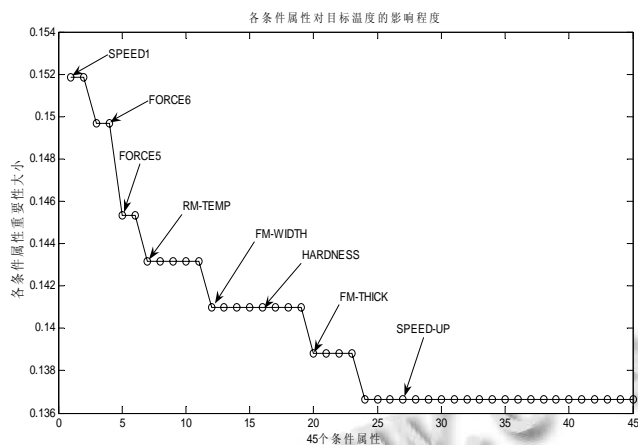


图1 各条件属性对目标温度的影响程度

在所有可能对热轧有影响的属性中,我们结合热轧工艺模型,选择属性重要性大的属性如:粗轧温度、带钢厚度、穿带速度和机架轧制力等作为分析对象,提取出对热轧目标温度影响大的属性。

粗轧出口处是一个条件较好的温度测量点,其测得的带钢表面温度可通过计算得到带钢的实际平均温度,是精轧温度计算的起始温度,可见粗轧出口温度对终轧温度是有很重要的影响。由实验结果可知,其粗轧出口温度对终轧温度的重要度为0.143167,且在五个属性中属性重要性非常大,也证明了其对终轧温度的影响非常重要,因此,粗轧出口温度可被选作为关键特征变量。

在温度控制模型中,与带钢厚度有关的压下率等都是温度控制模型中的重要参数;提高穿带速度可以减少带钢在机架间的辐射时间,以减少辐射造成的温降,且带钢厚度和穿带速度的属性重要度也非常大,因此,带钢厚度和穿带速度也可被选为关键特征变量。

通过特征变量提取系统后,提取的关键特征变量包括:粗轧出口温度,带钢厚度,穿带速度以及机架轧制力等,并将这些关键变量作为终轧温度控制模型

的输入,以提高终轧温度控制的精度,有力保证热轧带钢产品的质量。

4 结论

本文提出的基于属性重要性的热轧关键变量提取方法,是从整个决策系统的角度上对各个影响因素进行分析,提取出对热轧目标温度有影响的若干因素。处理的办法就是从整个样本集中去掉某些属性,看这些属性的删除是否会对样本的分类有影响,若删掉后分类的变化大,则其属性或是属性集在整个样本集中的重要性高。确定各条件属性对于目标属性的重要性大小后,再结合属性重要性大小和热轧工艺特点,提取特征属性。从实验中表明该算法在热轧工艺实践中是可行的,为建立可靠的终轧温度控制模型奠定了基础,可有效的提高精轧控制精度。

参考文献

- 1 Pawalk Z. Rough sets. International Journal of Computer and Information Science, 1982,11(5):341-356.
- 2 Li Y, Li FC, Jin CX, Feng T. A rule extraction algorithm based on attribute importance. Proc. of the Eighth International Conference on Machine Learning and Cybernetics, 2009,12(15),127-132.
- 3 汉德.数据挖掘原理.北京:机械工业出版社,2003.
- 4 Mitchell T.机器学习.曾华军,张银奎译.北京:机械工业出版社,2003.
- 5 Pawlak Z, et al. Rough set theory and its application to data analysis. Cybernetics and Systems, 1998,9:661-668.
- 6 杨成福,舒兰.基于属性重要性的决策表属性约减算法.计算机技术与发展,2006,16(11):62-66.
- 7 Wang BY, Zhang SM. A Novel Attribute Reduction Algorithm Based on Rough Set and Information entropy theory. 2007 International Conference on Computational Intelligence and Security Workshops,2007,21(2):81-84.
- 8 龚殿尧,徐建忠,刘相华,王国栋.热连轧带钢终轧温度的影响因素.东北大学学报(自然科学版),2006,27(7):763-767.