doi: 10.3969/j.issn.1005-3026.2016.12.009

基于大数据的C-Mn钢数据预处理及神经网络模型

吴思炜, 曹光明, 周晓光, 刘振宇
(东北大学 轧制技术及连轧自动化国家重点实验室, 辽宁 沈阳 110819)

摘 要: 在神经网络建模时,如果原始数据不加处理或经过简单剔除异常值后用于建模,则可能建立出错误的模型,即其规律并不符合物理冶金原理. 因此建模前需要对原始数据进行处理,使其呈现出显著的规律性. 针对钢铁生产采集的大量C-Mn钢数据进行了钢种归并,提出了数据预处理的一套方法,并采用LM-BP神经网络建立了满足一定精度(94.21%)的多牌号C-Mn钢屈服强度预测模型. 通过平均影响值(mean impact value,MIV)分析了成分及工艺参数对屈服强度的影响规律. 结果表明,随着碳含量的增加,屈服强度增大;随着终轧厚度和卷取温度的降低,屈服强度增大.

关 键 词: 大数据;建模;预处理;平均影响值;C-Mn钢
中图分类号: TG 335.11 **文献标志码**: A **文章编号**: 1005-3026(2016)12-1710-06

Data Preprocessing and Neural Network Model of C-Mn Steel Based on Big Data

WU Si-wei, *CAO Guang-ming*, *ZHOU Xiao-guang*, *LIU Zhen-yu*
(State Key Laboratory of Rolling and Automation, Northeastern University, Shenyang 110819, China. Corresponding author: LIU Zhen-yu, E-mail: zyliu@mail.neu.edu.cn)

Abstract: In neural network modeling, it may build a wrong model using original data without any treatment or only eliminating the abnormal value, for it could contain the law not to follow the physical metallurgy principle. To make the regularity significant, the original data need to be processed before modeling. In this work, based on the data of the C-Mn steel derived from a large number of data collected from different steel grades, a set of method for data preprocessing was proposed and a model for predicting yield strength of the C-Mn steel was established using LM-BP neural network, which could make the prediction accuracy meet the requirement (94.21%). The effects of the elements content and processing parameters on the yield strength were analyzed by the mean impact value (MIV). The results showed that the yield strength increased with the increase of carbon content and increased with the decrease of final rolling thickness and coiling temperature.

Key words: big data; modeling; data preprocessing; mean impact value (MIV); C-Mn steel

近年来,智能制造的提出加快了我国两化融合的进程,与此同时,通信、大数据及云计算等技术得到了迅猛发展,这些技术改变了传统的钢铁行业生产方式,同时也催生出了钢铁行业内的新技术. 其中,以力学性能预测为基础的集约化生产技术得到了较大发展. 采用生产数据建立力学性能预测模型,在一定范围内,采用控轧控冷技术,使用同一种化学成分的板坯制造出不同强度级别和用途的产品. 目前在采用大数据建模的研究中[1-5],都是将数据直接用来建模,而神经网络训

收稿日期: 2015-07-28
基金项目: 钢铁联合基金重点项目(U1460204); 辽宁省自然科学基金资助项目(2015020180); 中央高校基本科研业务费专项资金资助项目(N140704002).
作者简介: 吴思炜(1989-),男,辽宁阜新人,东北大学博士研究生; 刘振宇(1967-),男,内蒙古赤峰人,东北大学教授,博士生导师.

练数据的数据预处理的过程没有得到充分的重视. 文献[6]将关注点放在模型精度的预测，忽视了对模型规律性的研究. 如果深入研究模型中输出变量随输入变量的变化曲线则会发现不符合物理冶金规律的现象，在利用该模型对工艺进行反向优化时可能会产生错误的结果. 产生这种现象的原因是钢铁生产工艺的波动和性能检测的随机误差. 原始生产数据中混杂着较多的异常数据，这些异常数据使得原始数据规律性不够显著，进而影响所建立模型的合理性. 除此之外，相似工艺条件下的大量生产数据存在过多的重复信息，如果将过多的含有重复信息的数据用于建模，会加大建模的计算量，因此需要从大量生产数据中提取出含有重要信息的数据，去除冗余数据.

本文结合钢铁生产工艺的特点，针对以上问题，对钢铁工业大数据的预处理方法进行了探索，以多种牌号的 C－Mn 钢数据为例进行数据预处理和建模，根据所建立的模型分析了各影响因素对屈服强度的影响.

1　基本理论

1.1　分层聚类[7]

在钢铁工业的大数据中，需要选出工艺相近的数据，将工艺相近的数据进行归并. 设工艺参数分别为 X_1, X_2, X_3, X_4, X_5，计算 5 个参数间的马氏距离，则由分层聚类可得图 1，根据需要可以选择合适的分割点，将数据分成不同组类.

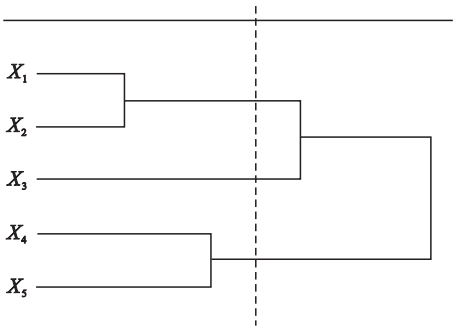

图 1　分层聚类示意图

Fig. 1　Hierarchical clustering profile

1.2　异常值的剔除

设某一炉钢卷的生产数据

$$\boldsymbol{P} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1i} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2i} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mi} & \cdots & X_{mn} \end{bmatrix}. \quad (1)$$

其中：n 为变量个数；m 为钢卷卷数.

不妨设钢卷数据中屈服强度为 X_{j1}，计算 m 组数据的中位数 M，均值 μ 和均方差 σ.

如果 $m < 30$，则采用改进的格拉布斯(Grubbs)法剔除异常值. 计算每卷钢的屈服强度剩余误差绝对值 $|V_j| = |X_{j1} - M|$，选择绝对值最大的一组数据，求出 G 值：

$$G = \frac{|X_{j1} - M|}{\sigma} = \frac{|V_j|}{\sigma}. \quad (2)$$

设置显著性水平为 α，则对照格拉布斯临界值 $G_{(n,\alpha)}$ 表查出数据个数为 n 时的格拉布斯临界值 $G_{(n,\alpha)}$. 比较 G 与 $G_{(n,\alpha)}$，如果 $G > G_{(n,\alpha)}$，则对应的第 j 组钢卷数据为异常数据，将其剔除. 将剩余的钢卷数据重复以上过程，直到没有异常数据.

如果 $m > 30$，则每一卷钢所对应的屈服强度会呈现正态分布，这时可采用拉依达(Pauta)准则. 若对于某一钢卷的屈服强度剩余误差 $V_j = X_{j1} - \mu$，有

$$|V_j| = |X_{j1} - \mu| > 3\sigma. \quad (3)$$

即屈服强度 $X_{j1} \notin [\mu - 3\sigma, \mu - 3\sigma]$，则认为这卷钢的屈服强度为异常数据，并予以剔除. 将剩余的钢卷数据重复以上过程进行处理，直到没有异常数据.

1.3　数据平滑

剔除异常数据后，将余下的数据平整化求均值，消除过多的包含重复信息的冗余数据，使每一炉钢保留一组稳定有效的数据.

$$\bar{X}_i = \sum_{j=1}^{m'} X_{ji}/m'. \quad (4)$$

其中，m' 为剩余数据数目. 结果为

$$\boldsymbol{P}' = [\bar{X}_1 \quad \bar{X}_2 \quad \cdots \quad \bar{X}_i \quad \cdots \quad \bar{X}_n].$$

1.4　神经网络平均影响值

平均影响值(MIV)是衡量神经网络中输入神经元对输出神经元的影响的一个指标，其符号代表相关性的正负，绝对值大小代表影响的相对重要性[8]. 在神经网络训练完成后，将训练数据 $\boldsymbol{P}$ 中每一个输入神经元在其原值基础上分别加/减 10% 构成两个新的训练数据集 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$，将 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 分别作为测试数据进行预测，得到预测结果 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$，求得 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$ 的差值后按照样本数求其平均值，即为该输入神经元 MIV.

2　数据预处理方法

钢铁工业大数据的预处理主要分为四部分：选择数据样本、填补空缺值、钢卷归并和相似工艺聚类，其流程如图 2 所示.

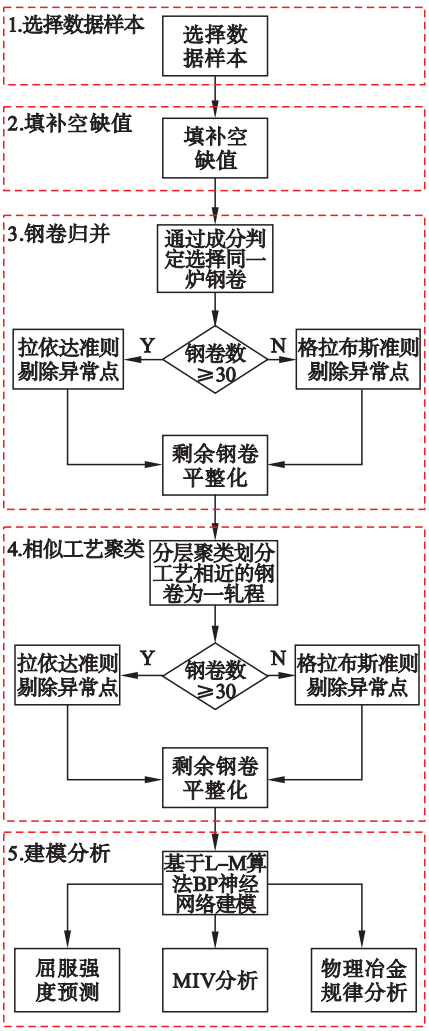

图 2 流程图

Fig. 2 Flow chart

本文以某钢厂生产的 C – Mn 钢为例进行数据预处理. 依据选择同一类别不同强度级别钢种建模的原则,采用多种牌号钢的工业生产数据作为数据样本. 剔除原始数据中记录不完整的数据,剩下完整数据共 6 454 组. 根据每条钢卷数据的主要成分判定其是否为同一炉钢,按照钢卷归并原则剔除异常值,对剔除异常值后的钢卷数据进行平整处理. 图 3 为某一炉生产的 12 卷钢的屈服强度分布,根据改进的格拉布斯准则判断得知第 6,7 卷钢数据(365 和 355 MPa)为异常值,故将其剔除. 对剩下 10 组钢卷的生产数据求平均值,得到屈服强度为 343 MPa,能够反映这一炉钢在特定生产工艺下屈服强度的平均水平.

钢铁生产工艺制定有着自身的特殊性,工艺的制定是离散的,并且所检测到的力学性能会有比较大的浮动,这两个特点确定了需要对工艺参数相近的数据进行归并. 通常,所采集到的数据会有三种情况:第一种为生产工艺参数在制定的工艺标准范围内,但是检测到的力学性能存在较大的偏差. 由于生产工艺比较稳定,因此检测到的力学性能在统计结果上其数值是比较集中的,呈现近似正态分布,如图 4 所示. 力学性能产生偏差的数据均为小概率事件,为增强数据规律性,将小概率异常部分剔除,此外将稳定的工艺参数和集中的力学性能用平均值表征这一工艺参数下所呈现的物理冶金规律. 第二种情况为生产工艺参数不在制定的工艺标准范围内,但是检测到的力学性能比较准确,符合物理冶金原理;无论生产工艺参数是否在制定的工艺标准范围内,由于这一部分生产工艺参数及力学性能对应的数据是符合物理冶金原理的,因此对于建立模型有利的信息,必须加以保留. 第三种情况为生产工艺参数不在制定的工艺标准范围内,这一类数据出现概率很小,对模型精度影响不大,为了保证模型的规律性允许其存留在数据中. 鉴于以上三种情况,本文对 C 含量、Si 含量、Mn 含量、终轧厚度(FDH)和卷取温度(CT)三种成分和两个轧制工艺参数进行分层聚类,使得每一类的成分、工艺参数相近,其成

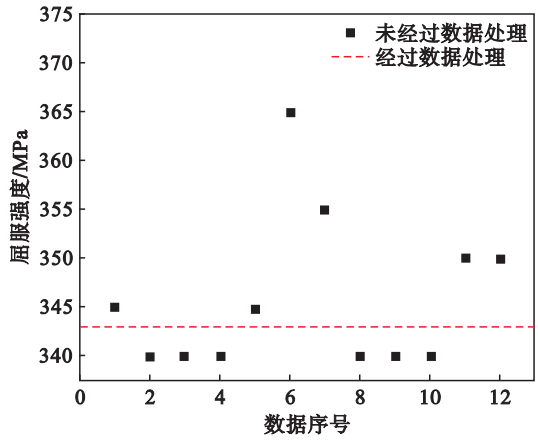

图 3 钢卷归并

Fig. 3 Steel rolls merging

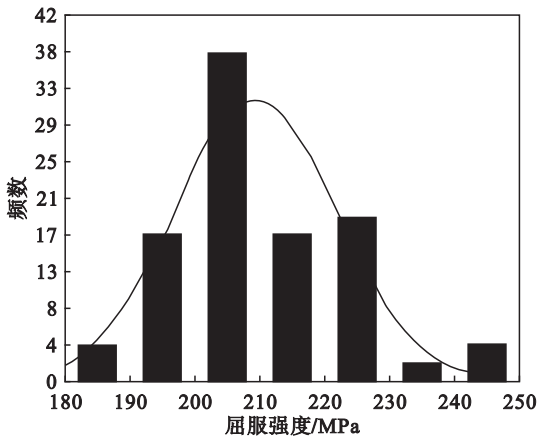

图 4 某一相似工艺下的钢卷屈服强度分布

Fig. 4 Yield strength distribution of steel rolls in a similar process

分和工艺参数数值控制在 $w_C \leqslant 0.02\%$，$w_{Si} \leqslant 0.15\%$，$w_{Mn} \leqslant 0.06\%$，FDH ≤ 0.8 mm，CT ≤ 20 ℃. 完成工艺聚类后，分析每一工艺类别力学性能数据分布情况，根据数据分布情况的不同选择不同数据处理方案. 如果力学性能数据呈现出标准的正态分布，则采用拉依达准则剔除异常数据. 如果数据较少，不符合正态分布，则采用改进的格拉布斯准则剔除异常数据. 最后将每一类别剩余数据用一组平均数据代替.

采用以上方法完成对所有数据的处理，最终得到具有代表性的数据 606 组. 相比原始数据，处理后的数据在数量上有了很大的精简，少量且具有代表性的数据可以减少建模的运算量，同时由于去除了冗余数据，处理后的数据具有更显著的规律性. 图 5 为在一组力学性能递增的轧制工艺下的原始屈服强度数据和处理后屈服强度数据的分布. 在原始数据中，工业生产条件的波动和力学检测的误差导致了数据规律性的模糊. 例如 1，2，3 组，4，5 组以及 9，10，11 组工艺下的屈服强度数据在统计上规律性不够显著，甚至在局部产生错误的规律. 经过数据处理后，数据呈现出稳定且显著的规律.

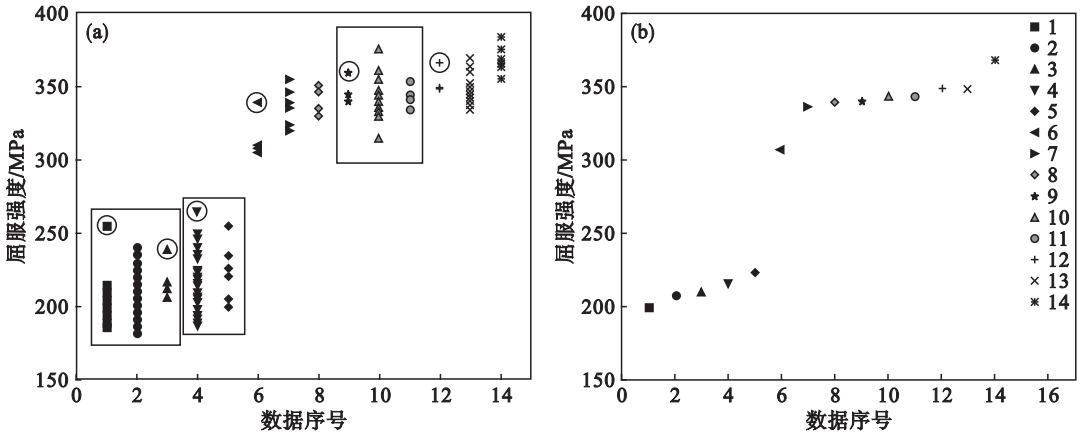

图 5　一组力学性能递增的轧制工艺下的屈服强度数据分布

Fig. 5　The distribution of yield strength data in a set of rolling process with increasing mechanical properties

(a)—数据预处理前；(b)—数据预处理后.

3　神经网络建模

神经网络建模采用基于 Levenberg－Marquardt 算法进行优化的 BP 网络，一个隐藏层，8 个隐藏神经元，分别选取 w_C，w_{Si}，w_{Mn}，中间坯厚度(FEH)、粗轧出口温度(RDT)、终轧厚度和卷取温度作为输入神经元，屈服强度作为输出神经元. 建立神经网络模型，并根据模型计算各工艺参数的 MIV.

为了比较数据预处理对数据建模的影响，分别基于未经过数据预处理的 6 454 组和经过数据预处理的 606 组数据进行建模. 将数据按照约4∶1分为训练数据和测试数据两部分. 未经过数据预处理和经过数据预处理的测试数据分别命名为测试数据 1 和测试数据 2. 分别将未经过数据预处理和经过数据预处理所建立的模型命名为模型 1 和模型 2. 分析两个模型中各因素对屈服强度的影响.

表 1 为模型 1 和模型 2 的 MIV. 在模型 1 的 MIV 中，w_C 和 FEH 的 MIV 为负，即 w_C 和 FEH 与屈服强度值成负相关关系，而 w_{Si}，w_{Mn}，RDT，FDH 和 CT 的 MIV 为正，即 w_{Si}，w_{Mn}，RDT，FDH 和 CT 与屈服强度值成正相关关系，其中 w_C，FDH 和 CT 与屈服强度的关系并不符合物理冶金原理. 而在模型 2 的 MIV 中，w_C，w_{Si}，w_{Mn}和 RDT 的 MIV 为正，FEH，FDH 和 CT 的 MIV 为负，即 w_C，w_{Si}，w_{Mn}和 RDT 与屈服强度值成正相关关系，而 FEH，FDH 和 CT 与屈服强度成负相关关系，符合物理冶金原理. 产生这种现象的原因是未经过数据预处理的数据中存在较多的异常值和小范围波动的值，使屈服强度产生错误的对应关系，因此导致所建立模型的规律性与物理冶金原理不相符.

表 1　各输入神经元的 MIV

Table 1　MIV of the input neurons

输入神经元	MIV(模型 1)	MIV(模型 2)
w_C	－0.517 4	5.053 5
w_{Si}	1.191 5	0.696 6
w_{Mn}	5.787 0	9.796 8
FEH	－1.830 5	－2.373 5
RDT	1.520 7	0.254 2
FDH	3.318 3	－0.140 4
CT	3.684 1	－13.324 4

图 6 为模型预测的屈服强度随输入神经元变化曲线. 为了验证模型包含的对应关系,图 6 中的散点是在其他成分和工艺相近情况下选取不同 FDH 和 CT 的实际生产检测的屈服强度的数据. 图 6a 直观反映了 FDH 对屈服强度的影响. 当成分和其他工艺一定时,板坯的屈服强度随着 FDH 的增大而降低. 这是由于在生产中,当中间坯厚度相同时,小的 FDH 对应较大的精轧压下量. 大的压下量产生大量形变,提高了储能,因此形核率增加,再结晶奥氏体晶粒尺寸减小,同时大量的位错缠结增大位错开动的阻力,使屈服强度增大. 此外,FDH 越小,冷却速度越大,更易获得较小的铁素体晶粒尺寸,获得细晶强化. 模型 2 中屈服强度随着 FDH 的变化规律一致,而模型 1 中由于多种工艺参数交互作用产生了错误的拟合结果.

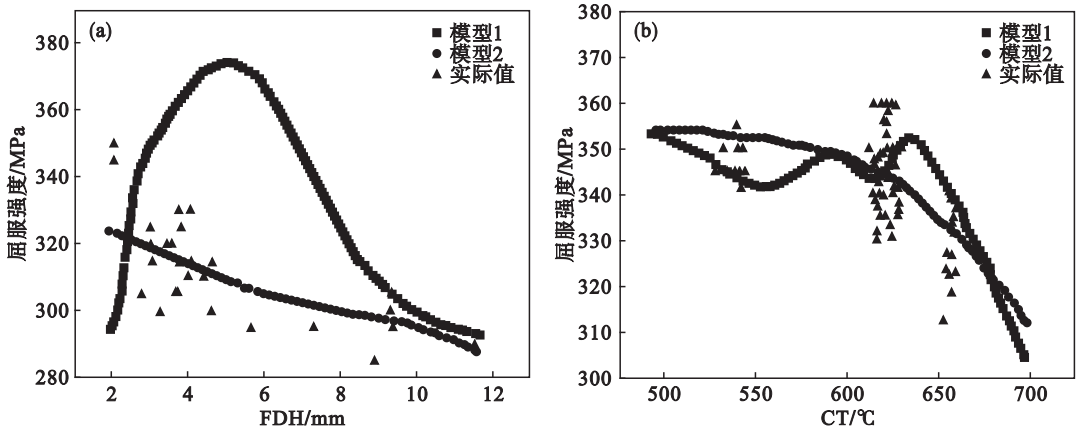

图 6 屈服强度随输入神经元变化曲线

Fig. 6 Curves of yield strength versus input neuron

(a)—FDH; (b)—CT.

CT 对屈服强度的影响如图 6b 所示,屈服强度随着 CT 的升高而降低. 当 CT 较高时,由于铁素体过冷度较低,形核点少且主要集中在原奥氏体晶粒的晶界处,铁素体晶粒长大较快,因此产生晶粒粗大均匀的铁素体. 当 CT 较低时,铁素体形核数目增多,生长速率降低,铁素体晶粒尺寸减小,同时珠光体呈弥散细小的状态分布. 随着 CT 的降低,铁素体晶粒尺寸减小,针状铁素体的数量增多,珠光体含量增多,其片层间距逐渐减小,因此,屈服强度增大[9-10]. 模型 1 也有随着 CT 的升高,屈服强度降低的趋势,但是 CT 在 550 ~ 650 ℃之间呈现起伏状,这种不稳定的状态是生产线采集的原始数据存在过多的异常值造成的.

表 2 为模型 1 和模型 2 对两组测试数据的预测结果. 精度度量采用预测值与实际值绝对误差在 ± 30 MPa 内的数据百分比. 采用测试数据 1 时,模型 1 的预测精度为 92. 66% ,模型 2 的精度为 89. 59% . 在模型 1 中,训练数据采用实际生产数据,数据包含着较多的误差,因此建立的模型包含随机误差. 而模型 2 采用剔除误差的数据,其预测精度较未剔除异常值有所降低,这是因为实际测试数据包含测量误差的结果,故模型 2 的预测精度低于模型 1. 采用测试数据 2 时,模型 1 的预测精度为 96. 25% ,模型 2 的预测精度为 94. 21% . 模型 1 的训练数据中存在大量的重复数据,而经过数据预处理的数据可以视作原始数据的子集,因此经过数据预处理的测试数据很大程度是包含在模型 1 的训练数据中的,因此其预测精度高于模型 2. 在将模型应用到智能制造的过程中,模型的合理性是智能系统优化出正确工艺的前提,必要时精度可以适当降低. 因此在建模时,力求保证一定精度前提下建立符合物理冶金原理的模型.

表 2 模型预测精度比较

Table 2 Comparison of predicted precision of models

%

模型	测试数据 1	测试数据 2
模型 1	92. 66	96. 25
模型 2	89. 59	94. 21

4 结 论

1) 提出了针对钢铁工业大数据的数据预处理方法,在保留原有特征信息的前提下,有效降低了数据的总量,去除了含有重复信息的冗余数据,使数据呈现出显著的规律性.

2) 在保证模型具有一定精度的前提下,建立

(下转第 1739 页)