文章编号: 1001-0920(2001)03-383-02

# 数据采掘在连铸板坯缺陷预报系统中的应用研究

张邦礼, 林 雪
(重庆大学 自动化学院, 重庆 400044)

摘 要: 利用数据采掘中的分类规则, 在 $k$-means 聚类方法中引入遗传算法, 进行连铸生产过程中板坯缺陷的诊断和预报, 得到了令人满意的结果。

关键词: 数据采掘; 遗传算法; 聚类; 连铸
中图分类号: TP 391 文献标识码: A

# Research of Data Mining in Continuous Casting Plate Slab Defect Forecasting System

*ZHANG Bang-li, LIN Xue*
(Institute of Automation, Chongqing University, Chongqing 400044, China)

**Abstract**: By means of clustering rules in data mining, the genetic algorithms were introduced into $k$-means clustering method. The diagnose and forecast for the plate slab defect in the process of continuous casting system were presented. A very satisfying result is concluted.

**Key words**: data mining; genetic algorithms; clustering; continuous casting

## 1 引 言

我国冶金行业普遍存在着这样的问题: 由于板坯的各种缺陷以及力学性能指标检验的滞后, 导致产品降级或改判, 造成巨大的浪费。例如, 连铸生产过程中产生的缺陷只能在轧制工艺之后才会发现。由于连铸的工艺参数复杂, 而且各个参数间相互关联, 人们对造成板坯缺陷的很多因素又不甚清楚, 因而单凭经验进行判断已远不能满足现代化、大批量、高质量的要求。目前, 许多利用计算机进行预报、模拟的系统应运而生。然而它们是在工艺过程机理的基础上建立的模型, 这种模型的适应性差, 移植困难, 而且要求对连铸的工艺、机理理解得非常透彻, 这在实际中很难办到。为此, 我们提出利用数据采掘技术进行连铸板坯缺陷预报的新方法。

数据采掘是指从大量的数据中采掘出隐含的、未知的、对决策有潜在价值的知识和规则。按采掘的知识类型可分为多种数据采掘。本文主要针对分类规则[1], 在 $k$-means 聚类方法[2] 中引入遗传算法, 利用已有的数据, 通过训练和学习得到缺陷的分类规则, 使在输入新数据时可以判定可能出现的缺陷及其程度, 克服了 $k$-means 算法收敛速度慢, 容易陷入局部极值的缺点。

收稿日期: 2000-10-11; 修回日期: 2000-12-07
基金项目: 国家教育部博士点基金项目(98-2000)
作者简介: 张邦礼(1941—), 女, 重庆人, 教授, 从事数据库中的知识发现、故障诊断的研究; 林雪(1976—), 女, 吉林九台人, 硕士生, 从事数据库中的知识发现研究。

## 2 原理及算法

### 2.1 $k$-means 聚类方法

给定一有限样本集 $X = \{x_1, x_2, \cdots, x_n\}$ ($x_1, x_2, \cdots, x_n$ 均为数据属性), 以及一个整数 $k$ ($k \leq n$), $k$-means 算法将 $X$ 分成 $k$ 个聚类 $Q_1, Q_2, \cdots, Q_k$, 并使每个聚类中所有值与该聚类中心距离的总和最小。

聚类分析实质上是一个全局最优问题。$k$-means 虽能有效地处理大数据, 但由于是随机选取初始聚类中心, 因此时常终止于局部最优解。为克服这一缺点, 我们分两步加以解决:

1) 引入遗传算法;

2) 在执行遗传 $k$-means 之前对数据进行预处理, 完成初始分类[3], 并在此基础上选取初始聚类中心。

### 2.2 初始分类

对数据进行聚类时应遵循如下规则: 在数据较为密集之处设置聚类中心; 距离中心较近的数据归为一类, 称为此聚类中心的覆盖圈。在进行初始聚类时, 首先选择彼此间距离最小的点作为聚类中心, 并将距其一定范围内的样本划入其范围圈; 然后从剩余样本中选择彼此间距离最小的点作为聚类中心, 并确定其范围圈。当继续选择聚类中心和进行归类时, 不再考虑已经处理过的样本; 彼此间距离较大的点也不聚在一起, 而是分别列为单独的一类。反复进行上述过程直至所有样本都被聚类为止。

### 2.3 遗传 $k$-means 算法

遗传算法是一类模拟自然过程, 主要是模拟生物界自然选择和自然遗传机制转化过程, 用以求解复杂问题的随机搜索算法。将遗传算法引入模糊聚类分析, 可以充分发挥其在处理组合优化问题方面所具有的优势。所有聚类中心组成问题的解, 以向量距离的加权平均和为评价函数, 在遗传交换和变异算子的作用下, 不断获得新的解集(即新的聚类中心集合), 从而使聚类分析导向全局最优。

本文采用扩展的遗传算法, 解的表示采用浮点数, 而不是二进制串。一个聚类中心向量表示为一个基因, 所有聚类中心组成染色体(即一个个体), 多个个体组成原始种群。迭代收敛结束后得到聚类结果, 每个类 $Q$ 对应有一个聚类中心 $Z$, 计算出每个类的最大半径 $R_j = \max(\| x_{ij} - z_i \|)$, 于是得到分类规则

$$\| X - Z_j \| \leq R_j \Rightarrow X \in C_j \qquad (1)$$

## 3 仿真结果

利用上述算法, 对某大型钢铁企业炼钢厂的 200 个样本数据进行仿真。每个样本含有 9 个分量, 包括碳含量、吹 Ar 前包温、结晶器水压、进出水温差等。各参数调整如表 1, 其中 $\epsilon$ 为设置误差, $m$ 为 $k$-means 算法常数, $P_c$ 为交叉概率, $P_m$ 为变异概率, $r$ 为范围圈关系。

表 1 参数调整值

| $\epsilon$ | $m$ | $P_c$ | $P_m$ | $r$ |
|---|---|---|---|---|
| 0.000 000 1 | 3 | 0.7 | 0.03 | 4.5 |

数据被分为两类, 与原始记录对照可知, 这两类主要代表合格和纵裂。对新的数据进行验证时, 效果较为令人满意, 可以判定此数据出现纵裂的程度。

## 4 结 语

本文利用一种快速有效的遗传聚类方法对连铸板坯数据进行验证, 得到了比较满意的结果。可以预见, 当数据量增大且更丰富时, 会得到更细致的聚类结果, 基于此法的数据采掘技术在连铸板坯缺陷预报中具有广阔的发展前景。同时也可看出, 本文提出的算法仅针对数值属性, 而对于非数值属性, 可采用 $k$-prototype 算法。

参考文献:

[1] 王实, 高文. 数据挖掘中的聚类方法[J]. 计算机科学, 2000, 27(4): 42-45.

[2] Selim S Z, Ismail M A. $K$-means-type algorithms: A generalized convergence theorem and characterization of local optimality [J]. IEEE Pattern Analysis and Machine Intelligence, 1984, 6(1): 81-87.

[3] 李聪, 张勇, 等. 一种新的聚类算法[J]. 模式识别与人工智能, 1999, 12(2): 205-208.