文章编号：1004－0609(2004)03－0494－05

铝电解控制中灰关联规则挖掘算法的应用①

刘业翔[1]，陈湘涛[1]，张更容[2]，李 劼[1]，邹 忠[1]
(1. 中南大学 冶金科学与工程学院，长沙 410083；2. 广西大学 数学系，南宁 530004)

摘 要：在事务拓扑空间的基础上，将灰色系统理论引入属性集的关联规则挖掘中，提出了一种新的适用于铝电解工业控制现场的灰关联度挖掘框架，并给出了基于该框架的灰关联规则挖掘算法，即 Gray_CTL 挖掘算法。将算法分解为两个小问题：1) 计算关于时间属性的灰关联度，这是算法的核心；2) 挖掘灰关联规则。以电解槽的热平衡数据挖掘为例，对某电解槽一个月的生产数据进行分析后，获得的灰关联规则说明该段时间内分子比、槽电压等因素对温度的影响程度较大。

关键词：铝电解；数据挖掘；灰关联规则；算子；Gray_CTL 算法
中图分类号：TP 311；TF 811.522 文献标识码：A

Application of mining algorithm based on gray association rule in aluminum electrolysis control

LIU Ye-xiang[1], CHEN Xiang-tao[1], ZHANG Geng-rong[2], LI Jie[1], ZOU Zhong[1]
(1. College of Metallurgical Science and Engineering,
Central South University, Changsha 410083, China;
2. Department of Mathematics, Guangxi University, Nanning 530004, China)

Abstract: The theory of gray system was brought into the mining of association rule about attribute sets, which based on the transaction topology space. Moreover, a new framework of mining using gray association degree was brought forward, and it can be used in aluminum electrolysis process control. The mining algorithm of gray association rule, viz. the mining algorithm of Gray_CTL, was described under the new framework. The algorithm was divided into two parts: the first is to calculate the gray association degree about time attribute and is the core of the algorithm; the second is to mine the gray association rule. An example of mining the gray association rule in the thermal equilibrium data was provided. After the analysis of production data coming from some electrolysis cell within a month, the obtained gray association rule indicates that the molecular ratio and cell voltage have greater effects on temperature of electrolyte than any other factors.

Key words: aluminum electrolysis; data mining; gray association rule; operator; Gray_CTL algorithms

数据挖掘(Data mining)，又称数据库中的知识发现(Knowledge discovery in database)，是指从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘有趣知识的过程[1]。对数据挖掘技术的研究，已引起国际人工智能、数据库、生物工程、数学等领域的专家和学者的广泛兴趣[2-8]，被认为是目前具有广泛应用前景的一个重要的研究课题。在事务数据库中挖掘关联规则是这一领域最活跃的一个研究分支，它由 Agrawal 等[9]首先提出。关联规则的挖掘，主要集中在对算法的研究方面，其理论基础是支持度—置信度框架，即“90% 的客户在购买面包的同时也会购买牛奶”。目前，著名的挖掘

① 收稿日期：2003－05－29；修订日期：2003－11－20
作者简介：刘业翔(1930－)，男，中国工程院院士，教授，博士生导师.
通讯作者：陈湘涛，博士研究生；电话：0731－8830474；E-mail: lbcxt@163.net

算法主要有 Apriori，DHP，PARTITION 等。

在铝电解生产过程中，人们采用了先进的计算机控制系统实现生产的管理和控制。在控制系统中，大量的反映电解槽当前工作状态的数据被采集起来用以获取电解槽工作时的各种槽况，分析其热平衡和物料平衡，控制其效应。这些采样数据是非交易型的事务数据，其数据的格式和属性是固定的，每隔一定的采样时间就会接收到一组数据，如槽子的工作电压、平均电压、针振、摆动、系列电压、系列电流、效应发生时刻、效应持续等；此外，还包括电解槽的各种工艺数据，如分子比、温度、两水平等。但在生产现场，人们却急需对这些采集到的海量数据进行各种信息的分析，挖掘其内在的关系，以求进一步分析其槽况，调整相关的控制参数，获取更好的控制效果；或者根据其历史数据，分析病槽的成因，判断其走势。由于这些控制系统的采样数据具有非交易型的、不变的事务模式，因此采用现有的挖掘算法所获得的支持—置信式关联规则没有任何实际的意义，不能指导生产。

为了弥补现有关联规则挖掘的不足，本文作者将灰系统理论[10-14]引入关联规则的挖掘中，提出了一种适合于铝电解工业控制现场的关联规则挖掘算法，即基于灰关联度框架的灰关联规则挖掘算法。这种灰关联度框架，采用几何关系和曲线间相似程度的量化比较分析方法进行构建，实现灰关联分析，进而获得基于时间属性的灰关联规则。

1　灰关联规则的相关概念

在本小节，作者首先建立了灰关联规则的两个基本概念：事务拓扑空间和属性集。在这两个概念的基础上，引入文献[10]中的灰关联度及其计算公式，提出了灰关联规则的相关概念；最后，在代数系统下研究适应灰关联规则挖掘问题的相关算子。

1.1　概念的定义

定义 1(事务拓扑空间)：设 D 是一个具有时间属性的非空事务集合，Γ 是 D 的一个子集族，并且具有以下的 3 个性质：

1) Φ，$D\in\Gamma$；

2) 若 A，$B\in\Gamma$，则 $A\cap B\in\Gamma$；

3) 若 $\Gamma_1\subset\Gamma$，则 $U_{A\in\Gamma_1}\in\Gamma$；

则称 D 为一个事务拓扑空间，简称为事务拓扑。

根据定义，在铝电解监控系统中存储两分钟报表的采集信息的事务数据库(或数据仓库)，其时间属性是槽控机的采样时间，显然是一个事务拓扑空间。记这个监控事务拓扑空间为 ora8。

事务拓扑空间是灰关联规则的概念基础，其他的所有定义都建立在事务拓扑空间之上。

定义 2(属性集)：设 D 是事务拓扑空间，在 D 中，需要考察的每个事务的比较因素的集合 Ω，称为事务拓扑空间 D 的属性集。

在 D 中，记时间属性为⊕，则⊕$\in\Omega$。

例如，在监控事务拓扑空间 ora8 中，属性集的元素包括采样时间、电压、电阻、平滑电阻、斜率、类斜等。

在考察事务拓扑空间的属性集的时候，往往有一个参考的标准，以确定其他的属性相对于参考标准的变化率。这个参考标准，笔者称之为主属性，其他待考察的属性称为非主属性。

在文献[10]中，提出了灰色关联四公理，即规范性、整体性、偶对对称性和接近性。在此基础上，笔者提出灰关联度的如下定义：

定义 3(灰关联度)：设 $X_0=\{\langle x_0(k),\oplus\rangle\mid k=1,2,\cdots,n\}$ 是 Ω 中关于时间⊕的主属性序列，$X_1=\{\langle x_1(k),\oplus\rangle\mid k=1,2,\cdots,n\}$，…，$X_m=\{\langle x_m(k),\oplus\rangle\mid k=1,2,\cdots,n\}$ 是非主属性序列，其中，时间⊕为属性集中需考察的某一给定的时间区间，序列中 n 的取值由⊕确定。给定实数 $\gamma(\langle x_0(k),\oplus\rangle,\langle x_i(k),\oplus\rangle)$，若实数 $\gamma(X_0(\oplus),X_i(\oplus))=\frac{1}{N}\sum_{k=1}^{n}\gamma(\langle x_0(k),\oplus\rangle,\langle x_i(k),\oplus\rangle)$ 满足灰色关联四公理，则称实数 $\gamma(X_0(\oplus),X_i(\oplus))$ 为非主属性 X_i 相对于主属性 X_0 在时间⊕的灰关联度，简称灰关联度，而实数 $\gamma(\langle x_0(k),\oplus\rangle,\langle x_i(k),\oplus\rangle)$ 则称为在时间⊕的灰关联系数。

灰关联度的实质是量化属性集中非主属性相对于主属性曲线间的几何差别。

事务拓扑空间 D 上关于不同的时间属性⊕，可以计算出相应时间区间的灰关联度，D 上关于时间属性⊕的所有的灰关联度的集合，构成 D 的灰关联度集。

定义 4(半序关系)：设 D 是事务拓扑空间，G 是 D 上关于时间属性⊕的灰关联度集，所谓 D 上的半序关系是指满足下列条件的一个关系$\leqslant G\times R$：

1) $\forall x\in G$，$x\leqslant x$(自反性)；

2) $\forall x$，$y\in G$，若 $x\leqslant y$，$y\leqslant x$，则 $x=y$(反对称性)；

3) $\forall x$，y，$z\in G$，若 $x\leqslant y$，$y\leqslant z$，则 $x\leqslant z$(传

递性)。

在半序关系≤中，蕴涵着关于时间属性⊕的一个假设，即他们具有相同的时间区间。

定义 5(灰关联规则)：设 D 是事务拓扑空间，Ω 是 D 的属性集。设 A，$B\in\Omega$，若属性 A、B 在时间⊕相对于主属性 C 的灰关联度 A_0，B_0 满足半序关系：$(A_0,\oplus)\leqslant(B_0,\oplus)$，则称 B 在时间⊕优于 A，记为 $A\prec B$。本文作者称 $A\prec B$ 是 D 在⊕的一条灰关联规则。

事务拓扑空间 D 上的所有灰关联规则组成的集合 R，称为 D 的灰关联规则集。这些灰关联规则具有时间的特性，表明在不同的时间段，事务对象的影响因素不一样。

在现实中，用户需要挖掘对象在不同时间段中处于主要矛盾的相关因素，以调整决策，争取获得最优的效果。例如，在 ora8 中，可以挖掘出这段时间影响电解槽的主要是两水平，而下一段时间的主要因素则为分子比。

1.2 灰关联度的计算

灰关联度的计算，是灰关联规则挖掘的核心问题，下面以定理的形式给出计算的公式。

定理 1：设 $X_0=\{\langle x_0(k),\oplus\rangle\mid k=1,2,\cdots,n\}$ 是 Ω 中关于时间⊕的主属性序列，$X_1=\{\langle x_1(k),\oplus\rangle\mid k=1,2,\cdots,n\}$，$\cdots$，$X_m=\{\langle x_m(k),\oplus\rangle\mid k=1,2,\cdots,n\}$ 是非主属性序列，其中，时间⊕为属性集中需考察的某一给定的时间区间，序列中 n 的取值由⊕确定。令

$$\gamma(\langle x_0(k),\oplus\rangle,\langle x_i(k),\oplus\rangle)=\frac{\min\limits_i\min\limits_k|\langle x_0(k),\oplus\rangle-\langle x_i(k),\oplus\rangle|+0.5\max\limits_i\max\limits_k|\langle x_0(k),\oplus\rangle-\langle x_i(k),\oplus\rangle|}{|\langle x_0(k),\oplus\rangle-\langle x_i(k),\oplus\rangle|+0.5\max\limits_i\max\limits_k|\langle x_0(k),\oplus\rangle-\langle x_i(k),\oplus\rangle|} \tag{1}$$

$$\gamma(X_0(\oplus),X_i(\oplus))=\frac{1}{N}\sum_{k=1}^{n}\gamma(\langle x_0(k),\oplus\rangle,\langle x_i(k),\oplus\rangle) \tag{2}$$

则 $\gamma(X_0(\oplus),X_i(\oplus))$ 满足灰色关联四公理，即 $\gamma(X_0(\oplus),X_i(\oplus))$ 是灰关联度，简记为 $\gamma_{0i}(\oplus)$。

定理 1 的证明比较简单，详细过程可以参考文献[10]。灰关联度 $\gamma_{0i}(\oplus)$ 满足规范性，即 $0<\gamma_{0i}(\oplus)\leqslant1$，说明任何非主属性一定和主属性相关。

在事务拓扑空间的属性集中，每个属性的计量单位各不相同，其数据的量纲也不一致。由于不同量纲之间的数据不便于比较，或者在比较时难以得到正确的结论，因此，在计算灰关联度时，需要先对数据进行无量纲化的处理。

1.3 灰关联算子

对数据进行无量纲化的处理方法有多种，常用的主要有：初值化、均值化和区间值化。对这几种方法，笔者在代数系统下，仅给出初值算子的定义。

定义 6(初值算子)：设 $X_i=\{\langle x_i(k),\oplus\rangle\mid k=1,2,\cdots,n\}$ 是 Ω 中关于时间⊕的属性序列，D_1 为序列算子，且：

$$X_iD_1=\{\langle x_i(k),\oplus\rangle d_1\mid k=1,2,\cdots,n\}=\{\langle x_i(k)d_1,\oplus\rangle\mid k=1,2,\cdots,n\}$$

式中 $\langle x_i(k)d_1,\oplus\rangle=\langle x_i(k)/x_1(k),\oplus\rangle$；$k=1,2,\cdots,n$

则称 D_1 为初值算子。在使用初值算子的时候，要求属性序列的第一个元素不能为 0。

这 3 种算子，一般不混合使用，而是根据挖掘的实际要求，选择其中的一种。如果系统具有稳定增长的趋势，可以选用初值算子；如果考察系统的周期性变化，可以考虑均值算子；一般的情况，则可以使用区间值算子。

2 灰关联规则挖掘算法

2.1 算法说明

根据前面的定义和定理，给定一个事务拓扑空间 D，挖掘其灰关联规则的算法问题可分解为两个小问题：1)计算待考察的非主属性关于时间属性⊕的灰关联度；2)挖掘灰关联规则。和 Apriori 等算法相类似，第一个小问题是挖掘算法的核心和关键，第二个小问题的解决方法比较简单，只需要根据灰关联度和时间计算其半序关系即可，其具体实现可以通过 VB6.0 等可视化开发工具进行设计。

在解决第一个小问题时，需要先进行数据的无量纲化(或数据的标准化)，在本文以初值化算子为例进行计算。

在第一个小问题中，时间区间的选择和确定比较重要。区间太长，比如选择一年作为考察的对象，由于前期影响的作用，淡化了当前的相互关系；而区间太短，比如选择某一天作为考察对象，则其影响的效果过于短暂，不能反映他们的真实状况。根据实际工业现场的经验，选择一个月比较合适。

2.2 算法描述

Gray_CTL 算法主要用来解决灰关联规则挖掘算法的第一个小问题。

```
算法：Gray_CTL
输入：database DB / * 监控数据库 * /
      start_time / * 开始时间 * /
      end_time / * 结束时间 * /
输出：gray_degree itemsets / * 灰关联度集 * /
get att_datasets from DB where datetime between start_time and end_time order by datetime / * 选择符合条件的属性序列集 * /
att_datasets=func_initial(att_datasets) / * 利用初值算子初值化，也可以利用其他算子 * /
s_count0=count of attributes in att_datasets / * 考察的属性集的个数 * /
s_count1=count of sequences in att_datasets / * 序列的个数 * /
proc_minmax(att_datasets, s_min, s_max) / * 求极限差值 s_min 和 s_max * /
for i=1 to s_count0
{s_sum=0
for j=0 to s_count1
{g_rel(i, j)=(s_min+0.5 * s_max)/(abs(att_datasets(0, j-1)-att_datasets(i-1, j-1))+0.5 * s_max)
s_sum=s_sum+g_rel(i, j)}
g(i)=s_sum/s_count1}
return g(i)
```

2.3 算法分析

Gray_CTL 算法对灰关联规则挖掘的第一个小问题进行了描述。由于在灰关联度框架中，灰关联规则具有时间属性，故 Gray_CTL 算法的增量更新问题，归结于时间属性的灰关联规则的挖掘。

在 Gray_CTL 算法中，不需要对数据库进行完全遍历(这主要由时间属性来决定)，但需要对满足条件的记录集进行三次遍历。在事务拓扑空间的采样数据比较大时，Gray_CTL 算法的效率比较高。

3 实验结果分析

在铝电解过程控制中，影响电解槽的因素比较多。对电解槽的热平衡研究，传统的方法通常通过选取一定的典型样本进行分析。本文的实验则从灰关联的角度，考察一定的时间属性，对这段时间内电解槽的所有数据进行挖掘，分析其特性。

在电解槽的热平衡挖掘过程中，以槽温作为主要参考因素，考察氟化铝添加量、阳极行程、日总下料量、日均工作电压、铝水平、电解质水平和分子比等非主属性，从而建立热平衡的挖掘模型，进行灰关联规则的挖掘。表 1 所示是某厂 401# 槽一段时间的实际工艺生产数据。

利用 Gray_CTL 算法，挖掘的非主属性对主属性的灰关联度分别为：0.765、0.676、0.816、0.992、0.979、0.916、0.997。因而，在时间(2002-11-28，2002-12-27)内，其灰关联规则为：阳极行程≺氟化铝添加量≺日总下料量≺电解质水平≺铝水平≺日均工作电压≺分子比。

根据挖掘得到的这条灰关联规则，说明这一段时间内的生产，影响 401# 槽槽温的主要因素是分子比、日均工作电压、铝水平和电解质水平等，从定量的角度分析了槽子的热平衡状况；根据专家的

表 1　某槽的实际工艺生产数据

Table 1　Real technical production data of some cell

Date	Temperature/℃	Addition of AlF_3/kg	Journey of anode/mm	Alumina feeding/kg	Average pot voltage/V	Metal level/cm	Bath level/cm	CR
2002-11-28	9450	3	22	1 763	4.212	23	18	1.24
2002-11-29	952	4	9	2 596	4.190	23	19	1.23
2002-11-30	952	5	19	2 072	4.170	23	19	1.23
2002-12-01	951	6	16	2 763	4.215	24	19	1.23
…	…	…	…	…	…	…	…	…
2002-12-25	953	7	11	2 423	4.162	22	22	1.25
2002-12-26	953	7	12	2 365	4.223	22	22	1.25
2002-12-27	953	6	10	1 786	4.233	22	22	1.25

经验，挖掘的结果和槽子的实际运行状况相符合。

在槽况热平衡专家诊断系统[15]中，神经网络的输入参数的权重根据样本学习而得到。这种样本的选择，从总体上反映了专家的知识和经验；结合挖掘的灰关联规则结果，将这几个影响因素的灰关联度参与输入权重的计算，进行槽况的诊断和分析，获得的实验结果更加令人满意。

4 结论

1）传统的关联规则挖掘算法，由于其基于支持度—置信度框架，不是很适合铝电解工业现场。作者将灰系统理论引入关联规则的挖掘，提出了一种新的框架：灰关联度框架，以解决现场的数据挖掘问题；提出了事务拓扑空间、灰关联规则集等有关概念。

2）详细描述了灰关联规则挖掘的 Gray_CTL 算法，给出其实现步骤，并进行了算法分析。

3）以铝电解槽的热平衡挖掘为例，其实验结果表明，这种灰关联规则挖掘算法比较适合于铝电解工业现场的数据分析。与现场考核管理相结合后，更有利于生产。

REFERENCES

[1] Han J W. Data mining techniques[A]. Proceedings of the 1996 ACM SIGMOD international conference on Management of Data [C]. Canada: ACM, 1996. 545.

[2] Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective [J]. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(6): 914 - 925.

[3] Fayyad U, Chaudhuri S, Bradley P. Data mining and its role in database systems[A]. Proceedings of the 26th VLDB Conference[C]. Cairo, Egypt: Morgan Kaufmanu, 2000. 63 - 124.

[4] Mitchell T M. Machine learning and data mining[J]. Communications of the ACM, 1999, 42(11): 30 - 36.

[5] Tsur S. Data mining in the bioinformatics domain[A]. Proceedings of the 26th VLDB Conference[C]. Cairo, Egypt: Morgan Kaufmann, 2000. 711 - 714.

[6] Riedell E, Faloutsos C, Ganger G R, et al. Data mining on an OLTP system (nearly) for free[A]. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data [C]. Dallas, Texas, USA: ACM, 2000. 13 - 27.

[7] DING Qin, Khan M, Roy A, et al. The P-tree algebra [A]. SAC 2002[C]. Madrid, Spain: ACM, 2002. 426 - 431.

[8] Eirinaki M, Vazirgiannis M. Web mining for web personalization[J]. ACM Transactions on Internet Technology, 2003, 3(1): 1 - 27.

[9] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[A]. Proceedings of the ACM SIGMOD International Conference on Management of Data [C]. Washington: ACM, 1993. 207 - 216.

[10] 刘思峰，郭天榜，党耀国，等. 灰色系统理论及应用[M]. 第二版. 北京：科学出版社，1999.
LIU Si-feng, GUO Tian-bang, DANG Yao-guo, et al. Theory and Applications of Grey Systems[M]. 2nd ed. Beijing: Science Press, 1999.

[11] 黄贯虹，黄伟健，方 刚，等. 大鹏湾优势藻引发赤潮的灰分析[J]. 生态学报，2002，22(6)：822 - 827.
HUANG Guan-hong, HUANG Wei-jian, FANG Gang, et al. Grey analysis of red tide produced by superior alga in Dapengwan Bay, South China Sea[J]. Acta Ecologica Sinica, 2002, 22(6): 822 - 827.

[12] 施国洪，姚冠新. 灰色系统理论在故障诊断决策中的应用[J]. 系统工程理论与实践，2001(4)：120 - 123.
SHI Guo-hong, YAO Guan-xin. Application of grey system theory in fault tree diagnosis decision [J]. Theory and Practice of Systems Engineering, 2001 (4): 120 - 123.

[13] 李如忠. 基于灰关联理论的流域生态环境评价[J]. 合肥工业大学学报(自然科学版)，2002，25(3)：464 - 467.
LI Ru-zhong. Eco-environmental assessment of basins based on grey associative theory[J]. Journal of Hefei University of Technology, 2002, 25(3): 464 - 467.

[14] 何 浏，赵 金，杨芷华. 模糊灰色理论在围岩类别评定中的应用[J]. 武汉水利电力大学(宜昌)学报，2000，22(4)：295 - 298.
HE Liu, ZHAO Jin, YANG Zhi-hua. Application of fuzzy grey theory to surrounding rock classification evaluation[J]. Journal of Univ of Hydr & Elec Eng/ Yichang, 2000, 22(4): 295 - 298.

[15] 李民军. 大型预焙铝电解槽模糊专家控制器及新颖热平衡控制模型的研究[D]. 长沙：中南工业大学，1999.
LI Ming-jun. On the Fuzzy Expert Controller and Heat Balance Control Models for Large-Scale Prebaked Aluminum Reduction Cells [D]. Changsha: Central South University of Technology, 1999.

（编辑 何学锋）