基于差别矩阵的属性约简算法及其应用

苏志同,李晋宏,林满山
SU Zhi-tong,LI Jin-hong,LIN Man-shan

北方工业大学 信息工程学院,北京 100144
College of Information Engineering,North China University of Technology,Beijing 100144,China
E-mail:suzhitong@ncut.edu.cn

SU Zhi-tong,LI Jin-hong,LIN Man-shan.Attribute reduction algorithm based on discernibility matrix and its application.Computer Engineering and Applications,2010,46(7):221-222.

Abstract: There are many correlated parameters in the production of aluminum electrolyser.How to select a part of parameters for analysis is important for aluminum electrolyser.An improved attribute reduction algorithm based on discernibility matrix is proposed.The advantage of the proposed algorithm is that the redundant elements in discernibility matrix are avoided.Experiments on real aluminum electrolyser data show that the algorithm is effective.

Key words: aluminum electrolyser;rough set;attribute reduction;discernibility matrix

摘 要: 铝电解过程中存在着各种相互影响的工艺参数,如何从中选择一部分参数进行分析,对铝电解生产有着重要的意义。提出了一种改进的基于差别矩阵的属性约简算法,避免了普通差别矩阵中的重复元素。用真实的铝电解生产数据对提出的算法进行了验证,效果良好。

关键词: 铝电解;粗糙集;属性约简;差别矩阵
DOI: 10.3778/j.issn.1002-8331.2010.07.067 **文章编号:** 1002-8331(2010)07-0221-02 **文献标识码:** A **中图分类号:** TP301

铝是一种重要的轻金属,工业应用十分广泛。铝主要来源于铝电解,铝电解生产投入大、耗能多,因此对铝电解过程的研究对增产降耗意义重大[1]。铝电解过程中的工艺参数比较多(在50个以上),且工艺参数之间存在着相互影响,如何从中选择一部分参数进行分析,对铝电解生产有着重要的意义。

粗糙集是一种新的处理不精确、不完全与不相容知识的数学理论[2-3]。在粗糙集理中,属性约简[4]是重要研究内容之一,在保险[5]和软件工程[6]等领域有着广泛的应用。用基于粗糙集的属性约简来研究铝电解生产是一种全新的尝试。

提出了一种改进的基于差别矩阵的属性约简算法,去掉了差别矩阵中不起作用的重复元素,并将该算法应用于约简铝电解生产中冗余的参数,取得了较好的效果。

1 粗糙集理论

定义 1 一个决策表定义为 $S=(U,R,V,f,d)$, 其中,$U=\{x_1,x_2,\cdots,x_n\}$是论域;$R=C\cup D$ 是属性集合;$C=\{c_1,c_2,\cdots,c_r\}$是条件属性集;$D\neq\varnothing$ 为决策属性集;$V=\cup_{a\in R}V_a$是属性的值域;$f:U\times C\rightarrow V$,$d:U\times D\rightarrow V$ 是信息函数。每一个属性子集 $P\subseteq R$ 决定了一个二元不可区分关系 $IND(P)$:

$$IND(P)=\{(x,y)\in U\times U|\forall a\in P,f(x,a)=f(y,a)\}$$

关系 $IND(P)$构成了 U 的一个划分,用 $U/IND(P)$表示,简记为 U/P,U/P 中的任何元素$[x]_P=\{y|\forall a\in P,f(x,a)=f(y,a)\}$称为等价类。

定义 2 在决策表 $S=(U,R,V,f,d)$ 中,$\forall A\in R,X\in U$,记 $U/A=\{A_1,A_2,\cdots,A_s\}$,则称 $A_(X)=\cup\{A_i|A_i\in U/A,A_i\subseteq X\}$为 X 关于 A 的下近似集。

定义 3 在决策表 $S=(U,R,V,f,d)$中,设 $U/D=\{D_1,D_2,\cdots,D_k\}$表示由决策属性集 D 对论域 U 的划分,$U/C=\{C_1,C_2,\cdots,C_m\}$表示由条件属性集 C 对论域 U 的划分,其中 $C_i(i=1,2,\cdots,m)$称为基本块,称 $POS_C(D)=\cup_{D_i\in U/D}C_(D_i)$为 C 关于 D 的正区域。

定义 4 在决策表 $S=(U,R,V,f,d)$中,差别矩阵 $\boldsymbol{M}=(m_{ij})$,其元素定义为:

$$m_{ij}=\begin{cases}\{c_k|c_k\in C,f(x_i,c_k)\neq f(x_j,c_k),d(x_i,D)\neq d(x_j,D)\}\\ \varnothing,\text{否则}\end{cases}$$

定义 5 由差别矩阵 $\boldsymbol{M}=(m_{ij})$导出的差别函数 $f(\boldsymbol{M})$定义为:$f(\boldsymbol{M})=\wedge(\vee m_{ij}),m_{ij}\neq\varnothing$。

定义 6 在决策表 $S=(U,R,V,f,d)$中,设 $\boldsymbol{M}=(m_{ij})$是差别矩阵,对 $\forall B\subseteq C$,若 B 满足:(1)$\forall\varnothing\neq m_{ij}\in\boldsymbol{M}$,有 $B\cap m_{ij}\neq\varnothing$;(2)$\forall b\in B,B-\{b\}$不满足(1),则称 B 是 C 相对于 D 的基于差

基金项目: 北京市教育委员会科技发展计划项目(No.KM200710009006);北京市属市管高等学校人才强教计划项目。
作者简介: 苏志同(1963-),男,副研究员,主要研究方向:数据库与数据挖掘;李晋宏(1965-),男,教授,主要研究方向:复杂工业生产数据挖掘、模糊专家系统;林满山(1965-),男,高级工程师,主要研究方向:数据仓库与数据挖掘。
收稿日期: 2008-09-08 **修回日期:** 2008-11-19

别矩阵的属性约简。所有 C 的属性约简的交称为 C 的核(简称核)。

2　改进的基于差别矩阵的属性约简算法

基于差别矩阵的属性约简算法的主要思想是:首先利用差别矩阵导出差别函数,然后求解差别函数的析取范式,该范式中的每一个析取项即为一个约简。其算法的优点在于直观,易于理解,且能够很容易地计算出核与所有约简。但这种算法也存在着不足之处,即差别矩阵中会出现大量的重复元素(或元素之间存在包含关系)。为解决这一问题,提出如下重复元素的定义。

定义 7　设 a、b 为差别矩阵中的两个元素,若 b 中的属性包含了 a 中的属性,则称 b 为 a 的重复元素。

定理 1　设 $\boldsymbol{M}$ 为差别矩阵,$\boldsymbol{M}'$ 是将 $\boldsymbol{M}$ 中的重复元素置空后得到的差别矩阵,则由 $\boldsymbol{M}$ 和 $\boldsymbol{M}'$ 所得到的属性约简是相同的。

证明　若 b 为 a 的重复元素,那么由定义 7 可知,在利用差别矩阵求差别函数时,元素与元素之间是合取关系。由吸收律可知,重复元素 b 在合取关系中不起作用,故将其置空不影响差别函数的值,因此重复元素置空前或后的差别矩阵所生成属性约简是相同的。

在此基础上,给出改进的基于差别矩阵的属性约简算法。

步骤 1　求出 $POS_C(D)$,$IND(D)$。

步骤 2　求出简化的析取式:

$m=1$;$d(1)=A_1 \vee A_2 \vee \cdots \vee A_{|C|}$;

for($i=1$;$i\leqslant|U|$;i++)

　for($j=i+1$;$j\leqslant|U|$;j++){

　　$m(j,i)=\varnothing$;

　　if(($i\in POS_C(D)$ AND $j\notin POS_C(D)$) OR (($i\notin POS_C(D)$ AND $j\in POS_C(D)$) OR (($i,j\notin POS_C(D)$) AND ((i,j)$\notin IND(D)$))) then

　　for($k=1$;$k\leqslant|C|$;k++)

　　if $A_k(j)\neq A_k(i)$ then

　　$m(j,i)=m(j,i)\vee A_k$;

　　flag=0;

　　for ($n=1$;$n\leqslant m$;n++)

　　if $m(j,i)\subset d(n)$ then

　　{$d(n)=m(j,i)$;*flag*=1;}

　　else if $d(n)\subseteq m(j,i)$ then

　　{$m(j,i)=\varnothing$;*break*;}

　　if ($n>m$) AND (*flag*=0) then

　　{$m=m+1$;$d(m)=m(j,i)$;}

}

步骤 3　将 i=1 到 m 做合取,得到简化差别函数,若 $d(i)$中的属性个数为 1,则为核中的元素。

步骤 4　利用简化差别函数得到约简。

算法说明:算法中 $m(j,i)$表示差别矩阵的元素;A_i 表示第 i 个条件属性;$A_i(j)$表示决策表中第 j 个对象在属性 A_i 上的取值;$d(n)$表示析取式;m 为析取式的个数;*flag* 为是否运用吸收律的标志。

3　在铝电解生产中的应用

实验数据来源于某铝厂 350 kA 铝电解控制系统数据库。此数据库包括了电解槽的各种生产数据:效应记录数据、异常炉数据、日报表数据以及各种测量数据等。选择其中的测量数据表,以出铝量为决策属性,人工选择了 6 个对出铝量影响最大的属性作为条件属性。构建的决策表如图 1 所示,其中 Djwd 代表电解温度;Djzsp 代表电解质水平;FeCnt 代表铁含量;Lsp 代表铝水平;PotNo 代表槽号;SiCnt 代表硅含量;AlCnt 代表出铝量。

DDate	Djwd	Djzsp	FeCnt	Lsp	PotNo	SiCnt	AlCnt
一月	974	230	.132	280	9345	.061	正常
一月	971	230	.132	265	9345	.061	很少
一月	974	230	.122	265	9345	.061	少
一月	970	230	.132	280	9345	.061	少
一月	971	230	.122	265	9345	.061	少
一月	974	230	.132	280	9345	.061	正常
一月	974	230	.113	320	9345	.061	正常
一月	974	230	.132	265	9345	.061	少
一月	970	230	.122	265	9345	.061	正常
一月	971	230	.099	320	9345	.061	正常
一月	974	230	.132	265	9345	.061	少
一月	974	230	.152	280	9345	.061	少
一月	970	230	.122	280	9345	.061	正常
一月	974	230	.122	280	9345	.061	少
一月	974	230	.109	280	9345	.061	正常
一月	974	230	.122	265	9345	.061	少
一月	974	230	.113	265	9345	.061	正常
一月	974	230	.109	280	9345	.061	正常
一月	974	230	.132	265	9345	.061	少
一月	974	230	.109	265	9345	.061	少
一月	970	230	.113	280	9345	.061	正常
一月	974	230	.122	320	9345	.061	少

图 1　铝电解数据决策表

对图 1 所示的决策表,用普通的基于差别矩阵的属性约简方法和该文提出的方法分别构建差别矩阵,所得到的结果分别如图 2 和图 3 所示。为了节省空间,略去了部分空白区域。

		FeCnt, Lsp	Djwd, FeCnt	
		Djwd	Djwd, Lsp	
		Djwd, Fe...	FeCnt	
FeCnt, Lsp	Lsp		Djwd, Lsp	FeCnt, Lsp
FeCnt, Lsp	FeCnt, Lsp		Djwd, Fe...	FeCnt, Lsp
		Lsp	Djwd	
Djwd, FeCnt	Djwd, FeCnt		Djwd, FeCnt	Djwd
Djwd, Fe...	Djwd, Fe...		FeCnt, Lsp	Djwd, Fe...
		Lsp	Djwd	
		FeCnt	Djwd, Fe...	
Djwd, Fe...	Djwd, Fe...		Djwd, Fe...	Djwd, Lsp
		FeCnt	Djwd, Fe...	
FeCnt, Lsp	FeCnt, Lsp		Djwd, Fe...	FeCnt, Lsp
		FeCnt, Lsp	Djwd, FeCnt	
FeCnt	FeCnt		Djwd, FeCnt	FeCnt
FeCnt, Lsp	FeCnt, Lsp		Djwd, Fe...	FeCnt, Lsp
		Lsp	Djwd	
		FeCnt, Lsp	Djwd, FeCnt	
Djwd, Fe...	Djwd, Fe...		Djwd, Fe...	Djwd, Fe...

图 2　使用普通方法得到的差别矩阵

FeCnt				
		Djwd		
Lsp				

图 3　使用提出的方法得到的差别矩阵

从图 2 和图 3 可以看出,改进的差别矩阵不但整体上比普通的差别矩阵要简单很多,而且矩阵的每一项也相对简单。两种方法得到的属性约简结果是相同的,都是铁含量(FeCnt)、铝水平(Lsp)和电解温度(Djwd),即槽号、日期、电解质水平和硅含量对决策属性出铝量(AlCnt)影响较小。需要说明的是,不同电解槽的性能之间存在着很大的差异,这里得出的结论只适用于选定的电解槽,是否具有良好的可推广性,还有待于进一步的验证。

4　结束语

文章的创新点:提出了一种改进的基于差别矩阵的属性约简算法。该算法从简化差别矩阵中的元素入手,提出了重复元素的概念,并说明了去掉重复元素的差别矩阵与普通的差别矩阵效果相同。将该算法应用于约简铝电解生产中冗余的参数,取得了较好的效果。

(下转 248 页)