第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·工程应用技术与实现· 文章编号：1000—3428(2007)03—0245—03 文献标识码：A 中图分类号：TP311.13

粗糙集数据挖掘技术在丙酮精制中的应用研究

焦 锴，王 雄，熊智华
(清华大学自动化系，北京 100084)

摘 要：将基于粗糙集理论的数据挖掘方法应用于丙酮精制过程产品质量的预报。针对流程工业数据高维、构成复杂、连续性强等特点，改进了基本的粗糙集数据挖掘算法，并与模糊聚类等技术相结合，提出了一种适用于流程工业数据的粗糙集数据挖掘方法。在采用实际丙酮精制生产数据作为样本的实验中应用效果良好，表明该方法具有一定的实用价值。
关键词：粗糙集；数据挖掘；丙酮精制；产品质量预报

Research on Application of Data Mining Based on Rough Set Theory in Acetone Refining

JIAO Kai, WANG Xiong, XIONG Zhihua
(Department of Automation, Tsinghua University, Beijing 100084)

【Abstract】 Data mining based on rough set theory in process industry is applied to prediction of product qualities. The original data mining model based on the rough set is improved to be suitable for process industry data that would be of large dimension, complexity and continuous in time. Combined with FCM, a method of data mining technique based on the rough set theory in process industry is designed. The method obtains good performance in the experiment on the real industrial data of a practical acetone refining process.
【Key words】 Rough set; Data mining; Acetone refining; Product prediction

数据挖掘是从存放在数据库、数据仓库等信息库的大量数据中挖掘有用知识的过程[1]。数据挖掘主要应用于商业、金融等领域，而在工业领域，特别是流程工业领域中，由于数据信息具有海量、高维、数据构成变化复杂、连续性强等特点，因此数据挖掘应用具有相当难度。目前在国内外工业领域中，基于Apriori性质的关联规则挖掘，以及依据人工神经网络等理论的数据挖掘算法，在工艺参数优化、故障诊断等方面已取得良好的效果。新兴的粗糙集数据挖掘，也凭借其自身的特点，得到了越来越多的研究与应用。

1 粗糙集理论在数据挖掘中的应用

粗糙集理论是Z. Pawlak于 1982 年提出的一种数据分析理论[2]。随着该理论在数据决策与分析、模式识别、机器学习与知识发现等方面的成功应用，目前已成为信息科学最为活跃的研究领域之一。

1.1 粗糙集理论基本概念

有限集 $U \neq \phi$ 上的一个划分 $C = \{X_1, X_2, ..., X_n\}$，其中 $X_i \subseteq U$，$X_i \neq \phi$，$X_i \cap X_j = \phi$，$\bigcup_{i=1}^{n} X_i = U$，$i, j = 1, 2, ..., n$ 且 $i \neq j$。U 上的一族划分称为关于 U 的一个知识库 $K = (U, \mathbf{R})$，其中 $\mathbf{R}$ 为一族表示划分的等价关系。令 $R \in \mathbf{R}$ 为 U 上一个等价关系，$[x]_R$ 为包含元素 $x \in U$ 的 R 等价类。取 $\mathbf{P} \subseteq \mathbf{R}$ 且 $\mathbf{P} \neq \phi$，其中所有等价关系的交集 $\cap\mathbf{P}$ 也是一个等价关系，称为 $\mathbf{P}$ 上的不可分辨关系 ind(**P**)。当 X 无法表示为某些 R 基本概念的并集时，称 X 为 R 不可定义，R 不可定义集称为 R 粗糙集。

R 粗糙集 X 的下近似集为 $\underline{R}X = \{x \in U | [x]_R \subseteq X\}$，上近似集为 $\overline{R}X = \{x \in U | [x]_R \cap X \neq \phi\}$。$\mathrm{pos}_R(X) = \underline{R}X$ 为正域，$\mathrm{neg}_R(X) = U - \overline{R}X$ 为负域，$\mathrm{bn}_R(X) = \overline{R}X - \underline{R}X$ 为边界域。$\mu_R(x, X) = card([x]_R \cap X) / card([x]_R)$ 为 x 对 X 的粗糙隶属函数。

如果不存在ind(**R**)=ind(**R**-{R})，称 R 为**R**中必要的。如果每个 R 都为必要的，称 $\mathbf{R}$ 为独立的。取 $\mathbf{Q} \subseteq \mathbf{P}$，如果**Q**是独立的，ind(**Q**)=ind(**P**)，称**Q**为 $\mathbf{P}$ 的一个约简。$\mathbf{P}$ 中所有必要关系组成的集合称为 $\mathbf{P}$ 的核core(**P**)[3]。约简和核是知识约简中的基本概念。知识约简即在保持知识库分类能力不变的条件下，删除其中不相关或不重要的知识，是粗糙集理论的核心内容之一，也是粗糙集数据挖掘的理论依据。

1.2 粗糙集数据挖掘算法

在数据挖掘应用中，根据已有知识对问题论域进行划分，并对每个划分确定其对某个概念的支持程度，即肯定支持、肯定不支持和可能支持。在粗糙集理论中分别用正域、负域和边界域 3 个近似集合表示。采用知识表达系统和决策系统描述问题，可将粗糙集数据挖掘算法建立在一种直观的二维表的基础上[4]。

一个知识表达系统记为 $S = (U, A, V, f)$。其中 U 为论域，A 为对象属性的非空有限集合，$V = \bigcup_{a \in A} V_a$，$V_a$ 为属性 a 的值域，$f: U \times A \to V$ 是一个信息函数，为对象每个属性赋予一个信息值，即 $\forall a \in A$，$x \in U$，$f(x, a) \in V_a$。

$S = (U, A, V, f)$ 简记为 $S = (U, A)$。其分辨矩阵定义为一个

基金项目：国家自然科学基金资助项目(60404012)
作者简介：焦 锴(1982－)，男，硕士生，主研方向：数据挖掘在流程工业中的应用；王 雄，教授、博导；熊智华，副教授
收稿日期：2006-02-27 **E-mail：**sscat00@mails.tsinghua.edu.cn

$n\times n$ 矩阵，n 为对象包含元素数，矩阵元素 $a(x,y)=\{a\in A|\,f(x,a)\neq f(y,a)\}$，$S$ 的分辨函数定义为布尔函数 $\Delta=\prod\sum a(x,y)$。

记 S 中，$A=C\cup D$，$C\cap D=\phi$，C 为条件属性集，D 为决策属性集，具有条件属性和决策属性的知识表达系统称为决策系统。

目前常用的基本粗糙集数据挖掘算法可归纳为[4]：

(1)根据样本数据定义一个决策系统 $S=(U,C\cup D)$。

(2)通过分辨矩阵计算所有约简，作为初始节点，依次从各节点中去掉一个属性得到后继节点直到节点为空，根据所包含属性数对节点归类分层。

(3)对由节点包含属性的不可分辨关系所划分的不可分辨类计算相对约简，得到简化的一组规则。

(4)对粗糙隶属函数大于置信度的规则进行保留，用于对新的数据进行推理和决策。

粗糙集数据挖掘算法能够选取数据项主元，接受决策先验知识，具有较强容错与抗干扰能力，计算效率高，便于进行并行处理，具有很强的实用性和灵活性。

2 丙酮产品质量预报

粗糙集数据挖掘在流程工业领域中的应用还处于探索阶段。本文以某大型石化企业中的丙酮精制工艺参数和产品质量数据为例，对粗糙集数据挖掘在丙酮精制中的应用进行研究。

2.1 丙酮精制工艺流程

丙酮精制工艺流程如图 1 所示。

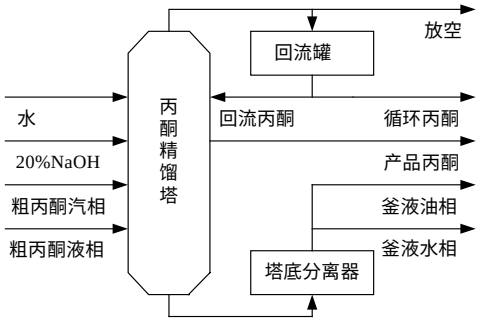

图 1 丙酮精制工艺流程

从上游装置粗丙酮塔中分离出的含丙酮、水和烃类的塔顶馏出物分离后进入丙酮精馏塔。氢氧化钠稀溶液和水也进入该装置，通过减压精馏，从侧线分离出高纯度的丙酮产品，同时从塔底脱除水、异丙苯及其它组分。回流罐中丙酮大部分回流至丙酮精馏塔，其余小部分采出至循环丙酮罐[5]。

丙酮产品质量标准分优级品、一级品、合格品 3 档，关键性评价参数为丙酮纯度和高锰酸钾褪色时间。由于目前产品质量抽样检测时间间隔太长，难以对产品质量实施监控，因此产品质量预报对于指导操作具有较大实际意义。

2.2 数据预处理和离散化

在工业领域，粗糙集数据挖掘应用首先遇到的难题就是工业数据的海量、高维、多时标性、不完备性和多模态性等因素的影响。在进行数据挖掘前，需要排除其干扰。

选取两批不同年份的实际生产数据分别进行实验分析。所选取时间段内丙酮生产装置运行稳定，数据记录完整。

丙酮精制生产工艺参数达 30 余种，根据工艺机理分析和现场生产经验，选取与丙酮产品质量有最直接关联的回流量、产品丙酮采出量、塔顶采出量、塔顶压力、塔釜温度、塔顶温度等 6 种工艺参数作为条件属性[5]。选取产品丙酮纯度和高锰酸钾褪色时间作为决策属性，参照优级品标准分为两类。6 种工艺参数数据采样间隔均为 1s，产品质量数据采样间隔为 3h。考虑到从粗丙酮进料到产品采出约半小时的流程时间，以及产品质量数据手工采样在时间点上的误差，选取产品质量数据时标为基准，对每一基准时间点之前 30min到之后 5min范围内的工艺参数数据求平均值，统一时标。每批数据预处理后共 160 组数据，140 组作为训练样本，20 组作为测试样本。

粗糙集数据挖掘算法所处理的是离散数据，需要对工业连续数据进行离散化。现在较常用的连续数据离散化方法是FCM模糊聚类算法，最初由Jim Bezdek提出，广泛应用于机器学习与数据挖掘中[6]。该算法根据对象之间的相似程度，将对象集合聚合成有限的几个类别，给出各数据的离散值，并返回各模糊区域聚类中心与各条件属性值对聚类中心的隶属度。该算法对于消除数据中的噪声干扰也有一定作用，当新数据来自训练样本空间覆盖范围以外时，也可通过模糊判别计算其离散值。

将每批数据利用 FCM 算法对训练样本条件属性进行离散化，条件属性聚类数均选取为 6，根据得到的聚类中心对测试样本条件属性进行离散化分类。

2.3 基本粗糙集数据挖掘算法的改进

工业生产数据经过预处理和离散化后即可运行粗糙集数据挖掘算法：(1)将训练样本作为决策系统 $S=(U,C\cup D)$，U 为生产数据组成的论域，C 为 6 种工艺参数组成的条件属性集，D为决策属性集；(2)计算决策系统的分辨矩阵及其分辨函数得到所有约简，即决策系统的核。对于两批数据均得到包括 5 种属性的一组约简；(3)计算相对约简，得到决策规则；(4)对每一决策规则，判定是否保留。

最后将得到保留的决策规则集用于对测试样本进行质量预报，考察预报正确率。

经分析发现基本粗糙集数据挖掘算法还存在 3 个问题：(1)工业数据受不确定性因素干扰较大，应采取措施去除受干扰数据产生的错误决策规则；(2)对于预报结果不同的规则，判定是否保留的标准应该不同；(3)某一组数据可能满足多组规则，需进一步判定，基本粗糙集数据挖掘算法采用的判定算法较复杂[4]，可以简化。

因此本文对基本粗糙集数据挖掘算法进行了如下改进：

(1)在判定某一决策规则是否应保留时，除对其粗糙隶属函数是否大于所设定的置信度进行考察外，还对训练样本中满足该决策规则的数据组数是否超过所设定的门限进行考察。这样在一定程度上去除了部分受干扰数据产生的错误决策规则。

(2)出现频度与置信度的门限则根据预报结果的不同，分别设置不同的值。

(3)在置信度相差不大的情况下，直接对规则根据预报结果的不同进行计数，比较计数结果，对该组数据的质量预报进行判定。

2.4 粗糙集数据挖掘的结果分析

对于相同数据，将目前工业数据挖掘中较常用的含非线性环节的前向 BP 神经网络算法和 Apriori 算法的计算结果，与粗糙集算法的结果进行比较。

验证比较结果如表 1 和表 2 所示，均为对各算法中的参数优化之后得到的结果。在改进粗糙集算法中，根据规则所

预报质量高低设置不同置信度门限；对每一组数据提取出所有该组数据所满足的规则，将计数最多的规则预报结果判定为质量预报结果。

表 1 粗糙集算法与其它算法效果比较(第 1 批数据)

算法	规则频度	规则置信度	运行时间	正确样本数	正确率
前向 BP	—	—	2s	13	65%
Apriori	2	0.80	12s	11	55%
基本粗糙集	—		5s	18	90%
改进粗糙集	2	0.80(y) 0.90(n)	5s	18	90%

表 2 粗糙集算法与其他算法效果比较(第 2 批数据)

算法	规则频度	规则置信度	运行时间	正确样本数	正确率
前向 BP	—	—	2s	11	55%
Apriori	2	0.80	12s	9	45%
基本粗糙集	—		5s	15	75%
改进粗糙集	2	0.80(y) 0.90(n)	5s	18	90%

比较发现，粗糙集算法的预报正确率较高，改进后性能更加稳定。Apriori 算法的预报正确率较低，并且运行时间较长。前向 BP 算法的运行时间较短，但是预报正确率不如粗糙集算法。

经过分析，粗糙集算法优于 Apriori 算法的原因包括：

(1)Apriori 算法对各数据项权重不加区分的对待，在不同条件属性对决策属性的影响程度相差较大时会产生较大误差。粗糙集算法首先通过对约简和核的计算，对数据项主元进行了选取，并在之后计算相对约简的步骤中进一步突出了主元对于规则的影响，使得规则更加可靠。

(2)Apriori 算法利用 Apriori 性质来提高运行效率。粗糙集算法计算相对约简的原理在本质上与 Apriori 性质是一致的，而对约简和核的计算，降低了数据维度，进一步提高了运行效率。

粗糙集算法优于前向 BP 算法的原因为：

(1)粗糙集算法基于对训练数据的统计计算结果，有能力在一定程度上消除数据中不确定性因素的影响。而前向 BP 算法在训练迭代计算中对于所有数据不加区分地进行误差统计，无法消除数据中不确定性因素的影响。

(2)粗糙集算法正确率完全取决于数据本身的精度。而对于绝大多数不能满足凸集条件的数据集，前向 BP 算法在训练迭代计算中有可能由于先验知识不足，网络权向量初值选取不当，而导致训练数据误差陷入局部极小，无法得到误差全局极小化的最优网络，产生较大误差。

相比于第 1 批数据，第 2 批数据的工艺参数更加稳定，变化幅度更小。考虑到工业噪声、测量设备误差等影响，第 2 批数据中不确定性因素的影响比重更大，这也为数据挖掘工作带来了更大难度。比较表 1 和表 2 也可发现，Apriori 算法、前向 BP 算法以及基本粗糙集算法的预报正确率对于第 2 批数据都发生了一定下滑，只有改进的粗糙集算法保持了稳定。这是由于改进粗糙集算法不仅保持了基本粗糙集算法的优点与特长，还利用增加可控参数的方法，提高了规则的可靠性。通过对多批不同数据进行实验，发现与基本粗糙集算法相比，改进的粗糙集算法的预报正确率更加稳定。

3 结论

粗糙集数据挖掘在流程工业领域中的应用还处于起步阶段，通过数据分析可发现粗糙集算法在丙酮产品质量预报中正确率较高，改进后得到的结果更加稳定可靠，在与两种常用算法的比较中优势明显，对于指导生产实践，改进工艺参数具有一定指导意义。如果进一步依据粗糙集理论各种研究成果，利用更多工艺生产流程的先验知识对粗糙集数据挖掘算法进行更多改进，进而与其它数据挖掘技术相结合，将会取得更好的效果。

参考文献

1 胡运发. 数据与知识工程导论[M]. 北京: 清华大学出版社, 2003.

2 Pawlak Z. Why Rough Sets?[C]//Proceedings of the 5th IEEE International Conference, 1996, 2: 738-743.

3 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.

4 李永敏, 朱善君, 陈湘晖, 等. 基于粗糙集理论的数据挖掘模型[J]. 清华大学学报(自然科学版), 1999, 39(1): 110-113, 117.

5 仲喃喃. 数据挖掘在丙酮精制产品质量预报中的应用研究[D]. 北京: 清华大学, 2004.

6 Hathaway R J, Bezdek J C. Fuzzy c-means Clustering of Incomplete Data[J]. Systems, Man and Cybernetics, Part B, 2001, 31(5): 735-744.

（上接第 224 页）

```
        begin
                OldBitMapArray[i].Min:=OldBitMapArray[i-1].Max;
OldBitMapArray[i].Max:=OldBitMapArray[i].Min+OldBitMapArray[i]
.VryFInall;
            end;
          end;
        end;
```

2.2 随机函数产生转盘效应

使用轮盘赌求出应该传递的图块序号

```
for i:=0 to 99 do   //轮盘赌
    begin
      percent:=Random(10000)/10000;
      for j:=0 to 99 do
        begin
          if (percent>= OldBitMapArray[j].Min) and (percent<
OldBitMapArray[j].Max) then
            begin
              …
            End;
        End;
    end;
```

3 结论

本文思想应用在工业电视，尤其是定点监控中，效果很好，比全帧传输数据量要减少 50%以上，而且在远程 Internet 上速度变得更快。在设计增益矩阵时，根据实际情况可以手动设定数据字典或者训练 BP 网络获得 K 增益阵的数据字典，通过一定时间的学习，图像的传递就可以达到最优状态。

参考文献

1 易继锴, 侯媛彬. 智能控制技术[M]. 北京: 北京工业大学出版社, 1999.

2 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社, 1999.

3 程云鹏. 矩阵论[M]. 西安: 西北工业大学出版社, 2001.

4 吕风翥. C++语言程序设计[M]. 北京: 电子工业出版社, 2001.