第十七届全国数据库学术会议论文集

（技术报告篇）

中国计算机学会数据库专业委员会（CCF DBS）
河北·保定　二〇〇〇年十月

河北大学出版社

责任编辑:马　力
封面设计:王占梅
责任印刷:李晓敏

图书在版编目(CIP)数据

第十七届全国数据库学术会议论文集.技术报告篇/
—保定:河北大学出版社,2000.10
ISBN 7-81028-667-6

Ⅰ.第...　Ⅱ.数据库系统-学术会议-文集
Ⅳ.TP311.13-53

中国版本图书馆 CIP 数据核字(2000)第 50237 号

出版:河北大学出版社(保定市合作路 1 号)　　经销:全国新华书店
印制:河北新华印刷一厂　　规格:1/16(787mm×1092mm)
印张:19.25　　字数:550 千字　　印数:1～1000 册
版次:2000 年 10 月第 1 版　　印次:2000 年 10 月第 1 次

定价:80.00 元

第十七届全国数据库学术会议

组织机构名单

会议主办单位

中国计算机学会数据库专业委员会

会议承办单位

河北大学　　华北电力大学　　河北农业大学

大会主席

王　珊

程序委员会

主席:李建中　　李天柱

委员:(按字典序)

陈良弼　杜小勇　冯玉才　何守才　何新贵　贾　焰　乐嘉锦　李战怀　李昭原　刘启原
陆宏钧　罗晓沛　马玉书　马应章　孟小峰　瞿兆荣　邵佩英　沈钧毅　施伯乐　孙志辉
唐常杰　唐世渭　童　頫　王国仁　王海洋　王洪水　王能斌　徐洁磐　杨冬青　于　戈
岳丽华　张　霞　周傲英　周立柱　周龙骧　孟卫一　宋晓宇　YU. PHILIP　张彦春

审稿人员:(按字典序)

何守才　何新贵　贾　焰　乐嘉锦　黄锦辉　黄上腾　李建中　李天柱　李战怀　李昭原
罗晓沛　马玉书　孟小峰　邵佩英　沈钧毅　施伯乐　孙志辉　唐常杰　童　頫　王海洋
王洪水　王　珊　徐洁磐　杨冬青　岳丽华　张　霞　周傲英　周立柱　周龙骧　张师超

组织委员会

主　席:傅广生

副主席:李天柱　陈景辉

成　员:黄　炜　高丽敏　王　军　徐建民　袁　方　王翠茹　郝书珍

编辑委员会

李天柱　肖艳芹　任建利

致　　谢

第十七届全国数据库学术会议的召开得到以下单位的大力支持，在此表示衷心的感谢。

国际商业机器中国有限公司

IBM DB2

美国 ORACLE 中国有限公司

ORACLE®

SOFTWARE POWERS THE INTERNET℠

塞贝斯软件（中国）有限公司

SYBASE SOFTWARE (CHINA) CO., LTD.

SYBASE®

INFORMATION ANYWHERE™

英孚美软件（中国）有限公司

Informix Software(China) Co., Itd

Informix SOFTWARE

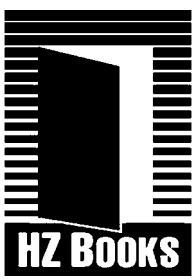
way to web

东方软件有限公司

NEUSOFT Corporation

北京华章图文信息有限公司/机械工业出版社

Huazhang Company/China Machine Press



计算机世界

南京南大谷元公司

山东地纬计算机软件有限公司

双狐软件有限公司

前　言

第十七届全国数据库学术会议(NDBC2000)于 2000 年 10 月 10 日在河北保定举行。会议由中国计算机学会数据库专业委员会主办,河北大学承办。自去年数据库专委会成立以来,在延续多年来形成的传统同时,积极寻求将这一传统的数据库盛会办成为聚集中国大陆、香港、台湾、澳门和海外华裔数据库专家学者交流学习的论坛,成为数据库研究人员、实践人员、应用开发人员、用户和企业交流有关数据库管理与应用的研究成果和实践经验,探讨今后数据库管理与应用所面临的关键性挑战性问题和研究方向的良好场所,并逐步使之成为亚太地区乃至世界性的有影响力的国际学术会议。美国计算机学会刊物 ACM Record No. 1, 2000 对数据库专委会的成立和本次学术会议(NDBC2000)作了报导。本次会议共收到论文 250 余篇,是历届会议较多的一次,论文来自海内外、高等院校、科研院所、企事业单位和公司,具有广泛的代表性。

2000 年 5 月在合肥召开了第十七届全国数据库学术会议的审稿会议。会议由中国计算机学会数据库专业委员会主任王珊教授主持。经大会程序委员会的认真评审,最后确定录用长文 108 篇,录用率为 43%,收入论文集 A 集(研究报告篇),由《计算机科学》设专辑出版;短文及会议交流 75 篇,收入论文集 B 集(技术报告篇),由河北大学出版社出版。本集为 B 集。论文基本上反映了我国目前在数据库领域的理论研究、实现技术及数据库应用等方向的研究成果,内容涉及了当今数据库研究的各个方面,如数据库理论、移动数据库、数据挖掘、数据仓库、半结构化数据与 XML、Web 数据库应用、海量数据存储、数据库安全、工作流管理等。既对当前世界研究热点领域的最新研究进行了跟踪,也对今后数据库研究和应用所面临的关键性挑战性问题作了探讨。反映了我国数据库界追踪国际前沿,为国民经济建设服务的研究水平与成果。

本论文集分为如下几个专题:

A. 数据仓库, 数据挖掘

B. Web 与数据库

C. 特种数据库

D. 数据库安全,数据库查询

E. 数据库应用

在论文集出版之际,我们对所有投稿者表示衷心感谢,向参加审稿的专家、教授致以深深的谢意。

我们特别感谢河北大学出版社在编辑出版本论文集中所付出的努力。

由于时间仓促,难免有缺点、错误,请不吝指正。

编者

2000 年 10 月

目　　录

A.　数据仓库，数据挖掘

B.　Web 与数据库

C. 特种数据库

D. 数据库安全,数据库查询

E. 数据库应用

数据挖掘在人事信息数据库的实现

周君毅[1]　毛勇锋[1]　许耀华[2]

(1 上海工程技术大学　上海　200336；2 上海市教育委员会　上海　200041)

Abstract　This paper presents a scheme of building three dimensional database structure based on data warehouse technology and the mechanism of Data Mining, which have implemented in the Personnel MIS for the Shanghai Institutions of Higher Education.

Key words　Data Mining　Database　Data Warehouse　DSS

1　引　　言

随着信息技术的发展,数据库在各行各业得到广泛应用。经过多年来的信息化建设,各机关和企事业单位都采集和积累了一大批原始数据和信息,这些数据和信息是宝贵的资源和财富,如何充分利用这些资源来为管理和决策服务,以提高决策的科学性,引起了管理部门和学术界的关注,而数据挖掘技术的发展为上述问题的解决带来了契机。

数据挖掘(Date Mining),亦称为数据库的知识发现,它是一个从大量数据中发现并提取隐藏在其中的、以前未知的、潜在有用信息和知识的过程。同传统数据库系统查询检索相比,数据挖掘是从系统内部自动获取隐含的、精练的知识的过程,它使用数据建立现实世界的模型,建模的结果是一种数据中的模式和关系的描述。

为了提高高校师资管理工作的科学化程度,充分利用各高校历年来所积累的有关师资情况的原始数据和统计报表等信息资源来为高校师资队伍建设的宏观决策和管理服务,上海市教育委员会有关部门和上海工程技术大学在科技攻关项目"上海市高校教职工信息系统"(简称为 PMSIHE)中研制了数据挖掘机制,在人事信息数据库应用数据挖掘技术方面进行了探索。

2　建立便于数据挖掘的数据存储体系

2.1　PMSIHE 的信息指标体系

PMSIHE 以关系模型的二维表为基础,采用开放式、动态的信息结构体系,即用户根据本单位管理和决策上的需要,可以随时通过系统提供的定义界面进行非基本信息数据项的增减及其数据结构的修改。为实现信息标准化,便于与其他人事信息系统实现数据共享,系统初始设置了现职教职工、单位情况、调离教职工和离休教职工四个信息群。各个信息群由一个基本信息集和若干个子信息集组成。各个信息群和信息集分别设置了一个数据表和对应的数据字典控制制表以及有关索引表。

2.2　应用数据仓库技术

对于只需使用当前数据的一般事务处理,上述的二维关系结构能满足其要求。而对于决策分析而言,由于要涉及大量的历史数据,许多决策方法必须通过对历史数据的分析的挖掘才能把握发展趋势,传统的关系数据库因自身的局限性而较难满足上述需求,因此人们研究和应用数据仓库技术来弥补关系数据库

的上述不足。

数据仓库是面向主题的、综合的、不同时间的、稳定的数据集合,用以支持管理中的决策过程。数据仓库除了传统数据库系统具有的共享性、完整性和数据独立性外,还具有面向主题、历史性和时间性等特征。由于数据仓库的数据具有时变特性,每个源数据都记录有时间标记,因此它能集成不同时间尺度上的数据,可以对历史信息进行管理。

为了实现对历史数据的处理和挖掘,笔者根据用户现有的计算机设备条件和信息资源的实际情况,应用数据仓库技术,通过采用具有时间标记的数据表文件命名规则以及数据字典映射和索引等机制,将原有二维的平面信息结构扩展为三维的立体存储结构,使其具有数据仓库的多维性和历史性特征。

2.3 三维的立体存储结构的实现

系统将各种相同类型的二维数据表(但时间标记不同)分别在相同方向进行叠加,从而组成一系列立体数据单元,使原来二维数据空间转变为具有时间属性的三维立体数据空间。每个数据单元具有惟一的代号 xxxx,表明其性质和类型,同时用于系统的引用和索引。

对每一个立体数据单元,系统分别建立了时间维映射表 T__xxxx 和字段维的数据字典 C__xxxx,如果数据表的性质属于交叉表类型,则系统为其建立一个记录维的数据字典 R__xxxx,这里 xxxx 为该数据单元的代号,立体数据单元的结构如图 1 所示。系统通过数据字典、映射表以及相关的索引实现对各数据单元和数据进行管理、连接和引用。

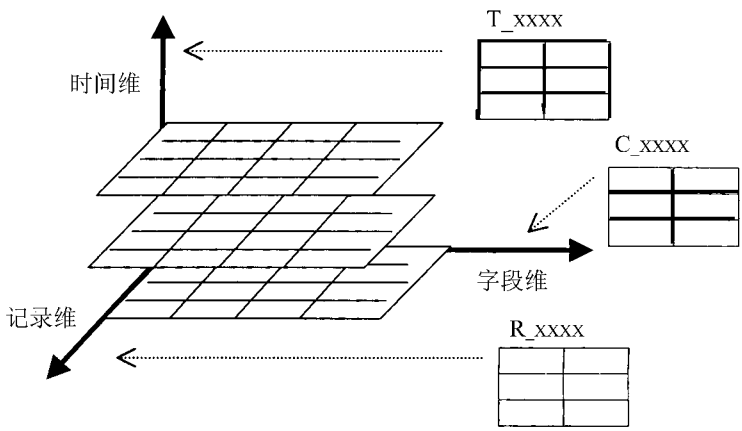


图 1 立体数据单元的结构

2.4 采用具有时间标记的表文件名命名方式

为使数据具有时间特征,根据人事信息变化频率和管理上的实际要求,系统把每半年作为一个数据采样的周期,即在每年 6 月和 12 月把各种原始数据表、各种报表和分析表保存起来,作为历史数据,不再更新。数据表文件名由类型域和时间域组成,采用以 xxxx__yyyymm 形式的 11 个字节的长文件名作为命名方式,其中类型域 xxxx 为该数据表的代号,表示该数据表的性质的类型,时间域 yyyymm 为时间标记,表明该数据表生成的日期,表中 yyyy 为年份,mm 为月份。例如:对于 1999 年 12 月统计的教职工基本情况表,则以 jb01__199912 作为文件名,这里 jb01 表示教职工表,199912 指明数据是在 1999 年 12 月生成。

2.5 数据时间标记的输入

系统具有定期把即时数据保存为历史数据的功能,并且自动追加数据的时间标记。对于已存在的无时间标记的历史档案和报表数据,系统则专门提供了历史数据导入工具,在导入历史数据时,系统根据用

户输入的数据生成时间自动追加时间标记。

3 数据挖掘的实现

3.1 数据挖掘的实现过程

PMSIHE 数据挖掘的实现是通过对数据源分别依次执行如图 2 所示的处理过程:系统对原始数据或分析报表等数据进行析取、切片和切块等操作,得到一次数据挖掘处理所需要的数据,接着选择挖掘工具进行挖掘和分析,然后把挖掘结果以报告或图表的形式输出。

3.2 自动和手工挖掘方式

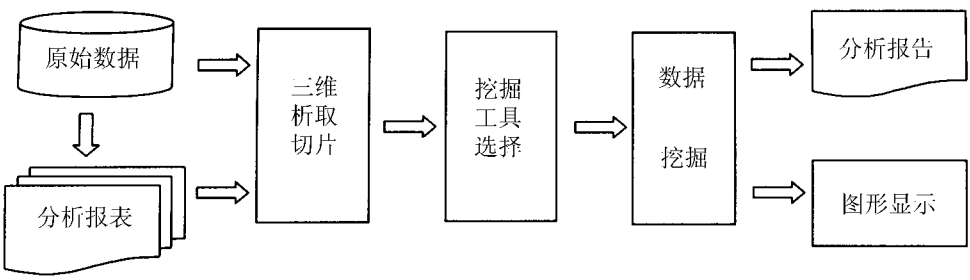


图 2 数据挖掘的实现过程

系统提供了自动和手工两种挖掘方式。自动方式是系统定期以批处理方式,依次在时间维、字段维和记录维的三个方向对数据源乾地搜索,按照将被调用的挖掘工具对数据域个数和形式的要求实施切片和切块,然后调用挖掘工具进行分析和计算处理,如果在挖掘中发现知识或找到数据中存在的关系和对应模式,则把这些挖掘结果的描述与相关数据的引用索引保存到挖掘结果数据库中,用于形成挖掘报告和以便用户日后查询。手工方式是用户根据需要进行定向挖掘。实现过程是:用户在界面上分别进行挖掘主题、挖掘工具、数据源及其范围、数据搜索方向与析取方式等的定义和选择,然后系统按照用户的要求进行处理,即时将挖掘结果反馈给用户。手工方式具有较大的灵活性和针对性以及时间响应快等特点。

3.3 自动切片(切块)的实现

为了简化,这里只讨论对某个数据单元进行字段维方向切块的过程。假定系统已保存 1985 年至 1999 年的教师职称与学历情况统计年报,报表标记为 jb05,该报表主栏(行)为教授、副教授、讲师等,宾栏(列)为研究生、本科、大专等。系统将这十五张表按年份依次叠加,组成如图 1 所示的数据单元。假设调用的挖掘工具是时间序列分析模型 tmol(),其数据接口为:变量 totl,放样本个数;二维数组 saml,放样本值。则数据切块的算法如下:

```
f=1
While f<=数据单元的字段个数                    && 字段维方向依次搜索切片
    r=1
    While r<=数据单元的记录个数                 && 记录维方向依次搜索切块
        i=1
        While i<=T__xxxx 表的记录个数           && 依次把该块的数据传送至数组 saml
          定位于表 T__xxxx 的第 i 条记录
          t=T__xxxx.时间标记
          b="jb05__"+str(t)                     && 形成数据表文件名
```

```
        定位于表(b)的第 r 条记录
        saml(i,1)=t                          样本的时间标记
        saml(i,2)=表(b)第 r 条记录的 Field(f)        && 样本值
      EndWhile
      totl=i                                 && 传送样本个数
      r=r+1
      Call tmol()                            && 调用数据挖掘模型程序
    EndWhile
    f=f+1
  EndWhile
```

对数据单元进行记录维方向切片的过程基本上同学段维方向进行切片的过程一样,而在数据单元的时间维方向进行切片相当于依次打开一个二维关系表,对其进行关系操作,所以这里不再赘述。

3.4 数据挖掘工具

数据挖掘工具采用相关分析、分类、回归、时间序列、分类归并和顺序发现等模型的算法。其中,分类和回归模型主要用于预测,相关分析和顺序发现模型主要用于描述或说明从用户数据库中捕获的行为,分类归并模型主要用于预测或说明。系统已提供相关分析、一元线形和非线形回归、多元线形回归、分类、时间序列等模型,其余模型正在进行研制。

通过数据挖掘,可以发现高校教师的学历层次越来越高,呈指曲线上升趋势;教师的年龄与科研成果有密切关系;教师的年龄与学历有较高的相关性以及科研成果与学科专业的分布关系等信息和知识。

4 结束语

本文介绍的数据挖掘实现技术和方案已经应用于上海市高校教职工信息系统,它能充分利用现有数据和信息资源,在较低的硬件配置下实现,具有实用性,对其他领域的数据分析和挖掘也有一定的参考价值。由于篇幅关系,关于数据挖掘工具的实现细节没能详细描述,同时该方案也存在某些局限性。数据挖掘是具有良好发展前景的研究领域,笔者将进一步深入该领域的研究。

参考文献

[1]Shapiro G P. From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining, 1966, pp. 1～35.

[2]姚宇蓉,何厚存,数据仓库中的信息存取分析技术的研究,计算机应用研究,1999,16 (8),54～58。

[3]周君毅等,上海市高校教职工信息系统的设计,上海工程技术大学学报,2000,14 (2),123～128。

基于粗糙集近似集扩展的规则提取算法*

卓 明 王丽珍 谭 旭

(云南大学计算机科学与工程系 99 研 昆明 650091 Master@zhuo.com.cn;
云南大学计算机科学与工程系导师 昆明 650091;
云南大学计算机科学与工程系 99 研 昆明 650091)

Abstract This paper presents a rule generation algorithm based on extending approximation in rough set , which can eliminate the noisy data by extending approximation and remove redundant data by pruning insignificant attributs with respect to the rate of coverage or reduction.

Key words Rough Sets Rule Generation Extending Approximation Rate of Coverage Rate of Reduction

1 引 言

1982 年波兰华沙理工大学的 Pawlak Z. 教授提出了粗糙集理论(Rough Sets)[5][6],这一理论对处理具有模糊、不精确或不完全信息的分类提供了一种新的工具。粗糙集通过数学公式计算得到,从而区别于依赖于统计方法的 Fuzzy 集。

数据挖掘是指从现实生活中的数据库中发现有趣或原来不知道的知识,可分以下几个部分[3]:数据选择、数据预处理、数据化简、综合评价。粗糙集的常用用途是在保持信息系统分类能力的基础上,进行知识约简。但考虑到样本的不完整性和噪音数据的影响,知识约简会受一定的限制,再说,仅仅找一个约简对一些问题来说是不是显得太局限了。因此本文立足于数据预处理和数据化简阶段,提出一种由粗糙集近似集扩展的算法来剔除噪音数据,及根据最低覆盖率?和最佳化简率选择属性进行约简,最后提取更为精简的规则来解决这一问题。

2 基于 Rough Sets 理论近似集扩展的规则提取算法

2.1 化简思想

传统的粗糙集理论中信息系统的约简问题用的是下近似集的概念。很容易想到,如果信息系统含有噪音数据,将会使求出的约简往往不是最简,且下近似集的要求比较高,不利于化简度的提高。文献[2][1]提出了一种用上近似集的方法分类和扩展信息表的方法,但此方法过于粗糙,且可能受属性次序的影响。本文综合了上述两种方法的优点,对两种方法的缺点加以改进并有所扩展,得到了以下算法。

首先,我们扩展了近似集,使判断的标准变为 $R*X\cap X\geqslant\lambda$($\lambda$ 为可信度,$\lambda\in[0.6,1]$),因为我们仅用下近似集分类时,要求覆盖所有数据,不利于信息的化简。当判断的标准变为 $R*X\cap X\geqslant\lambda$ 后,$R*X\cap X<\lambda$ 的部分往往是噪音,我们通过剔除 $R*X\cap X<\lambda$ 的信息自然也就消除了噪音。当然,当 $\lambda=1$ 时,其判断标准变成了下近似集判断。

其次,我们通过引入最低的覆盖率?(?$\in[0.7,1]$)来控制我们化简的程度,?指化简后的样本数量占

* 本文研究得到云南省自然科学基金资助。

原样本的最低限度。在考虑有噪音的影响下,我们可以适当放宽条件,使化简后信息消除了噪音,且包含的信息量在?之上。当然,当?=1时,问题就变成了信息的约简。

最后,为了排除因属性的排列次序的不同而造成化简结果的差异,我们提出化简率的定义:化简率 = 覆盖率 / 规则数 。即在相同覆盖率的前提下,选取规则数较少的方式化简。这样每次选取那些覆盖率尽量大而提取的规则尽量精简的化简,大大地提高了化简质量,且不因属性的排列次序而改变化简结果。

2.2 算法描述

输入:1>S=(U,A)是一个样本信息系统,U 为论域且 U={ X1,X2,……,Xn },A 是所有属性的集合。
CON 是条件属性集合,DEC 是决策属性集合,CON?DEC=A,
2>最低的覆盖率?, 样本数量 RECORD
3>可信度 λ

输出:输出化简后的样本信息系统和规则

```
1#    F=CON
2#    Threshold = ?                                  '最低的覆盖率
3#    For (j=|F|;j≥1;j - -)
4#      MinQuality = Threshold
5#      Reduct[1..j] = 0                             '置化简率数组为 0
6#      Found = False
7#        For (k=j;k≥1;k- -)
8#          F = F-C[k]
9#         CurrentQuality = | F * DEC∩DEC≥λ|/ RECORD '计算去除属性 k 后的覆盖率
10#         If (CurrentQuality ≥MinQuality) Then
11#           MinQuality = CurrentQuality
12#           Reduct[k] = CurrentQuality / | F * DEC | '计算去除属性 k 后的化简率
13#           Found=True
14#         End if
15#       F = F ⋃ C[k]                                 '恢复该属性
16#       End For
17#     If (Found == True) Then
18#       DelAttribute = Max { Reduct[1..j] }          '选择化简率最大的属性剔除
19#       F = F-DelAttribute
20#       C[DelAttribute]=C[j]                         '剩余属性合并
21#     Else
22#       Break                                        '跳出外循环
23#     End If
24#   End For
25#   剔除冗余属性
26#   剔除重复的记录样本
27#   输出化简后的样本信息系统
28#   提取规则
```

2.3 算法分析

算法中 18# 选择了满足最佳化简率的属性进行删除,这样保证了不因属性的排列次序而选择满足最低覆盖率,但化简率最高的属性进行剔除。9# 通过加入 λ 删除了噪音数据。10# 保证了规则的覆盖率。

该算法的时间复杂度是 O(l2 O(n2(l+m))),其中 l 是条件属性的个数,m 是决策属性的个数,n 是样

本的个数。O(n2(l+m))是计算 | F * DEC∩DEC≥λ| 的时间,总的来说,本算法的时间复杂度是多项式级的,并且可以根据检验的结果,再改变 ?、λ 的值进行相应的化简,直到满足需求为止。

2.4 算法举例

例:有如下气象学上的信息表

表 1

	CON					DEC
	Outlook	Temperature	Windy	Humidity	Pressure	Class
1	Rain	Cool	False	High	Middle	Bad
2	Overcast	Cool	True	High	Low	Bad
3	Overcast	Cool	True	Normal	Low	Bad
4	Sunny	Hot	False	High	High	Good
5	Sunny	Middle	True	High	High	Good
6	Rain	Middle	True	Normal	Low	Bad
7	Sunny	Hot	False	Low	Middle	Bad
8	Overcast	Cool	False	Normal	Low	Good
9	Overcast	Middle	True	Normal	Middle	Bad
10	Sunny	Cool	True	Normal	High	Bad
11	Sunny	Hot	False	Normal	High	Good
12	Rain	Middle	True	Low	Middle	Bad
13	Sunny	Hot	False	Normal	Middle	Good
14	Rain	Middle	True	High	Low	Bad
15	Rain	Cool	False	Low	Low	Bad
16	Overcast	Cool	False	High	High	Good
17	Sunny	Hot	True	High	Middle	Good
18	Overcast	Cool	False	High	Middle	Good

我们应用上述算法对此表进行化简:

1> 约简度 ?= 0.9, 可信度 λ=0.6182

去除属性 Pressure 后,CurrentQuality=1.0;去除属性 Humidity 后,CurrentQuality=0.944;发现噪音数据 7,化简后得到表 2。

提取规则如下:

a. (Outlook,Rain) ? (class,Bad)

b. (Outlook,Sunny) and (Windy,False) ? (class,Good)

c. (Outlook,Sunny) and (Temperature,Cool) ? (class,Bad)

d. (Outlook,Sunny) and (Temperature,Not Cool) ? (class,Good)

e. (Outlook,Overcast) and (Windy,True) ? (class,Bad)

f. (Outlook,Overcast) and (Windy,False) ? (class,Good)

2>约简度 ?= 0.8, 可信度 λ=0.618

去除属性 Pressure 后,CurrentQuality=1.0;去除属性 Humidity 后,CurrentQuality=0.944;去除属性 Temperature 后,CurrentQuality=0.889;发现噪音数据 7、10,化简后得到表 3。