

一种基于清洗关联规则的金融数据挖掘算法

刘丽峰

(山西金融职业学院, 山西 太原 030008)

摘 要: 充分研究新型的数据前期预处理及金融数据联系的一些时间变量的预处理方法, 提出一种基于清洗关联规则的金融数据挖掘算法. 引入关联规则的分析在一定支持度的约束下研究和金融紧密相关的一些因素, 在应用过程中随着算法的开展, 对重点部门与重点功能进行观测与预警. 实验证明, 该方法可以更加准确的进行金融数据状态的跟踪, 挖掘效率更高.

关键词: 金融系统; 关联规则; 清洗

中图分类号: TP391.9

文献标识码: A

文章编号: 1000-7180(2012)05-0174-04

A Cleaning Association Rules Based on the Financial Data Mining Algorithm

LIU Li-feng

(Shanxi Vocational Institute of Finance, Taiyuan 030008, China)

Abstract: The new type of the full data pretreatment and focuses on the research and financial data of some key contact time variable of the pretreatment method. Put forward a kind of cleaning association rules based on the financial data mining algorithm, introduce the analysis of association rules in certain support under the restriction of the research and financial closely related to some factors, in the application process with the development of key departments of system and the key functions of the observation and warning. Experiment results show that the method can be more accurate for the state of the financial data tracking, mining efficiency is higher.

Key words: financial system; association rules; cleaning

1 引言

计算机技术在银行的金融管理中得到了大量的普及与应用, 银行中的各个部门都建立的自己的金融信息管理系统, 如财务管理系统、结算管理系统、信息管理系统、账目信息管理系统、科研与试题库管理系统、数据管理系统等. 这些数据库与管理系统的开发不仅减轻了金融管理者的劳动强度, 提高了工作的效率, 而且还积累了大量的数据. 但是这些海量数据还没有得到有效的应用, 金融的管理者在日常的管理中只是通过简单的统计与排序筛选等潜在的方法获得表面的一些信息. 这些历史的庞大的数据都是来自日常工作与管理的实践, 它们中间蕴藏着丰富的信息就像一个带开发的“金库”一般^[1-5]. 如何

对这些表面的数据再次进行深层次的挖掘, 从这些信息中发掘更深层次的知识供决策者进行管理提高金融管理质量与管理水平是现在各大研究机构正在教学研究中研究的重点^[6-8].

2 金融数据清洗预处理

2.1 数据见表

目前很多银行的数据是否庞大, 甚至有几万多, 信息量的巨大已经足以引起管理者的重视, 纷纷创建了自己的数据仓库用来进行数据的展现与分析挖掘, 数据仓库是面向主题分析的一种大型数据库, 现在银行的数据仓库的数据源主要有两类^[9-10]. 一是各分行的金融管理系统, 它是从各分行的数据库中抽取出来, 如金融数据管理与信息管理. 另一类就

是平常的日常管理包括人力管理系统与各个场馆的管理与后勤的管理等,一个良好的数据仓库的完成设计体系应该改如图1所示.本文正是在某银行的数据仓库中进行了数据的抽样选取.在抽样的过程中采取多因素随机抽样的方法,将总体按照地域、金融条件、阶段、资本充足率等段等一些因素分类,然后在各分类中进行随机获得研究样本.

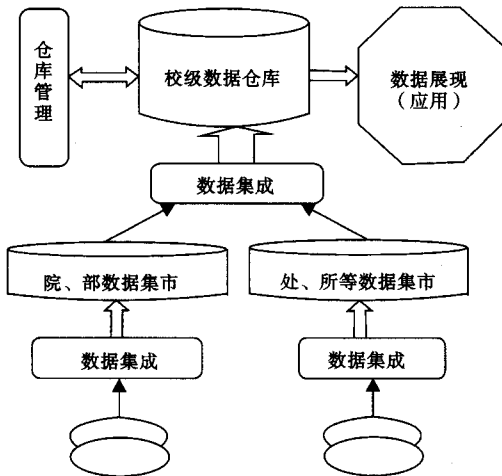


图1 银行系统数据仓库

在图1针对本文的研究主题学生课程学习成绩的主题数据库包括以下三部分:

(1) 初始数据库:初始数据库中,包含着一些银行操作的相关信息,例如用户姓名,身份信息,存款数量,存款以及取款的时间等等基本的信息.

(2) 基本数据库:基本数据库中的数据都是经过相关的预处理,以表格的形式存在的,形成真正的数据信息库,里面的数据已经具备了一定的关联性.数据之间存在数学关系..

(3) 数据挖掘库:这里面的数据是根据一定的数据挖掘算法处理后的数据,该数据可以之间通过相关的搜索算法被搜索到,数据中的信息也是为数据挖掘算法进行服务的.初始信息和基本数据库中的信息按照一定的关联规则,关联属性形成一个较为统一的整体,为数据的分类检索提供依据.

丰富的数据维数与细小的粒度划分是数据挖掘效果的重要影响因素,如果维数不全面与粒度划分不合理将会影响关联的效果,相反将数据进行过于细致的划分将会使分析的计算负担增强不利于系统的运行.

2.2 数据清洗规范处理

由于信息量较大,数据中可能存在对搜索起到副作用的冗余信息,这些信息的存在会大幅度的降

低数据之间的关联性,使得数据之间的联系降低,不利于挖掘算法的进一步操作,因此需要对数据进行清洗操作,最大程度去除冗余.

有些变量的由于某种原因的产生将可能很明显地失去了最小支持度与置信度的约束,比如在银行的管理过程中某银行因为事故等原因进行调整与长时间的数据不更新给金融数据分析带来的影响将应该被剔除,还有一部分变量如表1中的数据提交次数与出席活动的次数具有极大的相关性,变量的存在会给计算结果带来较为负面的作用,其中最为明显的是变量的存在会提供素数据的维度,一旦数据的维度增加,会增加算法的复杂程度,造成算法耗时.因此本文采用了以下的维数约简的方法.

(1) 利用众数分析,找出“低效”变量

通过以上分析,可以确定每个数据的众数是不同的,根据众数的不同,可以对数据和冗余进行有效的划分,最大程度的去处冗余数据的干扰,保证数据的真实性.

(2) 利用相关系数的分析,解决变量间的共线冗余.利用 Pearson 相关系数公式:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X}) - (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}) \sum_{i=1}^n Y_i - \bar{Y}}} \quad (1)$$

求解两两变量之间的相关系数,根据相关系数 $|r| > 0.8$ 的高度相关组中,选取一个作为代表.经过以上两种方法的约简可以大幅度的将一些冗余信息进行剔除.

2.3 变量的标准化处理

对以数据库中的数据,不同数据有着不同的属性,这些属性上的差异对不同的挖掘算法来说,作用不同,如果属性的差异过大,会影响挖掘算法的效率.比如说有的评价机构给银行的成绩为优、良、及格、不及格这样的等级制,而有的评价机构直接将银行的成绩进行百分制计数还有的评价机构将银行的成绩用5分制来进行计数.建立模型前应该对变量进行标准化或者归一化的处理使量纲的取值范围影响得到消除,本文使用Z分数标准化:把数据标准化为一个均值为0,标准差为1的Z分数.

$$X' = (X - \bar{X}) / S$$

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad (2)$$

可证:绝大多数的数据都会包含在 $[\bar{X} - 3S, \bar{X} + 3S]$ 范围内.

+3S]区间内,可以将区间以外的点作为孤立点剔除.

2.4 生成时间趋势属性

在银行的金融管理系统中,属性大致可以分为两类,一类是与时间无关的属性,包括货币的种类信息与持有人情况等等,一类是与时间相关的属性,包括金融对某一数据的缺勤趋势变化与数据更新上交情况变化等等属性.由于样本的数据收集是随着时间的进行而进行的,如果仅仅将某些属性进行简单的汇总无法反映该银行在管理过程中对它的兴趣与心态的变化,将直接影响管理的效率,因此需要对属性序列进行趋势分析生成趋势属性.

假设时间的记录粒度为月记为:

$$X = (1, 2, 3, 4, 5, 6)$$

出勤的课程总数记为 Y . 最小二乘法在线性回归中得到了广泛的应用,将上述的关系记为

$$Y = \alpha + \beta X \quad (3)$$

其中 α, β 为回归系数,分别表示直线在竖轴的截距与直线的斜率.可以使用最小二乘法求解使得数据与该直线的估计之间误差最小可用如下公式计算.

$$\beta = \frac{\sum_{i=1}^s (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^s (x_i - \bar{x})^2}$$

$$\alpha = \bar{y} - \beta \bar{x} \quad (4)$$

其中, $\bar{x}$ 是 x 的平均值,求得的回归系数 β 作为趋势属性取值.在进行建模的过程中出勤次数与月数之间的数值相差很大单位不同,为了使 β 值不至于发散,需要调整,本文解决的方法是先按照原方法计算 β 再使用下述的方法进行调整求得实际的回归系数 β' .

$$\beta' = \beta \frac{\bar{x}}{\bar{y}} \quad (5)$$

3 关联金融数据挖掘步骤

算法的第一步是对金融数据的有效挖掘,步骤如下:设 I 为 m 个项的集合, $I = \{i_1, i_2, \dots, i_m\}$; 交易 T 为项的集合,且 $T \subseteq I$; D 为 T 的集合. 对应的每一个数据处理都有一个相应的编号,编号可通过一些特殊的操作设定,设为 T_D . 设 X 由众多交易数据组成的集合,如果 $X \subseteq T$, 交易的具体信息在集合中能够找到. 包含的具体形式如: $X \Rightarrow Y$, 可有 $X \subset T, Y \subset T$, 而且 $X \cap Y = \emptyset$. 数据之间的关系系数表

示为: $X \Rightarrow Y$ 在数据集合 D 中的支持度(Support)是相关数据集中,同时包含 X 和 Y 的交易数与所有交易数之比,记为 $\text{Support}(X \Rightarrow Y)$, 即

$$\begin{aligned} \text{Support}(X \Rightarrow Y) \\ = \{T: X \cup Y \Rightarrow T, T \in D\} / D \times 100\% \\ = s\% \end{aligned} \quad (6)$$

规则 $X \Rightarrow Y$ 在交易集 D 中的可信度(Confidence)是指包含 X 和 Y 的交易数与包含 X 的交易数之比,记为 $\text{Confidence}(X \Rightarrow Y)$, 即

$$\begin{aligned} \text{Confidence}(X \Rightarrow Y) \\ = \{T: X \cup Y \subseteq T, T \in D\} / \{T: X \subseteq T, \\ T \in D\} \times 100\% = c\% \end{aligned} \quad (7)$$

以上公式还可以表现为如下形式:

$$\begin{aligned} \text{Confidence}(X \Rightarrow Y) \\ = \text{Support}(X \Rightarrow Y) / \text{Support}(X) \times 100\% \\ = c\% \end{aligned} \quad (8)$$

频繁项集是指支持度不低于最小支持度阈值的项集. 给定一个交易集 D , 挖掘关联规则是产生支持度和可信度分别大于用户给定的最小支持度(min-sup)和最小可信度(min-conf)的关联规则. 为了挖掘有意义的关联规则,需要给定 2 个阈值:(1)用户规定的关联规则必须满足的最小支持度,表示一组物品集在统计意义上需满足的最低程度;(2)用户规定的关联规则必须满足的最小可信度,反应关联规则的最低可靠度.

4 实验结果

为了验证本文提出算法的优势,进行算法对比实验,在模型的构造中进行了大量的数据实验来进行这个时间属性的变化研究,实验证明这种变化能更好的将是时间的属性因为数据的差异影响降到最小.

图 2 是四个银行在整个年度中的金融数据的进出率变化. 其中数据 1 与数据 2 的进出次数非常稳定,没有明显的波动,他们的状态在实际调查中,也显示的非常好. 若直接应用最小二乘法进行参数计算,得到的 β 值分别是 9.37 与 -1.657, 可以理解为数据 1 对该业务的业务掌握较为稳定,数据 2 存在着剧烈的变化有评价不及格的危险. 但是这与实际的情况是不符合,采用式(8)的调整以后这种状况得到了改善, β 值分别为 0.021 与 -0.009, 非常接近于 0 可以认为这两个数据业务能力的热情与认真程度是较为稳定的.

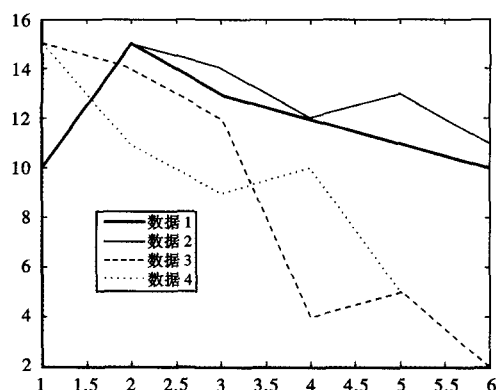


图2 某银行金融数据按月汇总情况

5 结束语

本文在数据清理后将模型基于 SPSS Clementine 数据挖掘平台,进行基于传统的 Apriori 算法的数据挖掘构架来进行某银行金融管理的数据挖掘.在进行实验的过程中反复调整支持度与可信度发现本文中提出的数据前期准备的一些列工作将对数据挖掘模型的有效性大大提高.将训练的模型结论与实际的情况进行比较发现该模型的关联准确率为 72%这个数字较大多数研究者的成果有了一定的提高.

参考文献:

[1] 李艳美,张卓奎.基于贝叶斯网络的数据挖掘方法

- [J]. 计算机仿真, 2008(02): 117-119.
- [2] 林锦贤,沈钧毅.教育管理数据仓库系统分析[J].福州大学学报:自然科学版,1999,27(6):14-17.
- [3] 许国英.论本科教学质量的控制[J].黑龙江高教研究,2004,23(5):83-84.
- [4] 张庆利,李文斐.浅谈提高本科教学质量[J].高教探索,2003,19(1):27-27.
- [5] 罗尧成,陈敬良,姚俭.我国高校课程与教学改革三十年:历程、经验与瞻望[J].中国高教研究,2009,24(2):11-14.
- [6] 狄浩林.基于成绩信息管理系统的高校教务管理研究[D].北京:北京邮电大学,2006
- [7] 杨王黎,吴雅娟,王丽侠.成绩分析与试卷质量评价系统的设计与实现[J].大庆石油学院学报,2002,26(2):60-6.
- [8] 冯亚丽,伊三泉,吴凯,等.基于WEB的成绩管理信息系统的安全性策略[J].大庆石油学院学报,2004,28(6):77-78.
- [9] 赵松,孙立铸. Apriori 算法的改进[J].哈尔滨理工大学学报,2007,29(5):94-97.
- [10] 曾波.一种基于单事务项集组合的频繁项集挖掘算法[J].计算机科学,2008,35(1):196-197.
- [11] 胡钢锋;李德毅;陈桂生;韩明畅.一种网络化数据挖掘方法研究[J].微电子学与计算机.2006.09.

作者介绍:

刘丽峰 女,(1970—),硕士,副教授.研究方向为人工智能与数据挖掘.

(上接第173页)

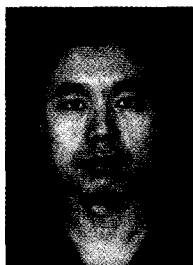
[9] 张盛意,蔡之华,占志刚.基于改进模拟退火的遗传算法求解 0-1 背包问题[J].微电子学与计算机,2011,28(2):61-64.

[10] 龚纯,王正林.精通 MATLAB 最优化计算[M].北京:电子工业出版社,2009.

作者简介:



杨久红 女,(1973—),工程师,硕士,副教授.研究方向为 BP 神经网络、智能优化算法.



王小增 男,(1973—),博士研究生,副教授.研究方向为机械结构疲劳、断裂、可靠性分析、力学性能分析.