DOI :10.11721/cqnuj20130221

一种改进的基于概念格的数据挖掘算法*

李志坚[1,2], 莫建麟[1,2]
(1. 电子科技大学 成都 610054 ;2. 阿坝师范高等专科学校 四川 汶川 623002)

摘要 :为解决多维数据模型与关系数据模型之间的双向数据系统查询、数据清洗、数据转换、实现集中和分发数据的准确性与一致性等问题,通过对概念格的相关研究,将全局数据挖掘与局部数据挖掘相结合,提出一种改进的基于局部信息的全局概念格的数据挖掘算法,并将挖掘过程分解为 ETL(Extraction-Transformation-Loading)动作,结合 ETL 处理工作流,实现并行分布式海量数据的时序挖掘。实验证明,该算法对增强数据加工能力具有一定的实用性。

关键词 :数据 ;概念格 ;数据挖掘 ;集成 ;时序挖掘
中图分类号 :TP301.6 **文献标志码** :A **文章编号** :1672-6693(2013)02-0092-04

概念格,又称 Galois 格[1]。自 20 世纪 80 年代 Wille 提出了形式概念分析以来[2],许多学者将之做为形式概念分析及表示工具进行了深入研究,并取得了许多重要成果。目前比较公认的对概念格的定义为:知识表示和知识处理的一种有效工具,它的基本思想是根据二元关系提出的一种概念层次结构,是数据分析和规则提取的一种有效工具[3]。

通过研究发现,目前有很多构造概念格的算法,国外对于概念格技术的应用相对比较成熟,主要集中于粗糙集、模糊集、本体、语义 Web 等一起研究。查阅文献对概念格的研究大致可分为两类[4] :一是渐进式算法,该类算法针对形式背景中第 i 个对象生成概念集或 Hasse 图 ;二是批处理算法,该类算法根据整个形式背景建造概念格及 Hasse 图。文献[5]详细分析了国内外对于形式概念分析的研究方向,如规则提取、属性约简、子格及商格、维护和概念格建格算法等基础理论研究,均属于渐进式算法的研究。而关于对海量数据进行批处理的算法目前由于受环境或自然条件的限制寥寥无几[6]。

为解决多维数据模型与关系数据模型之间的双向数据系统查询,数据清洗、数据转换等问题,本研究通过对概念格的相关研究,将全局数据挖掘与局部数据挖掘相结合,提出了一种基于局部信息的全局概念格的数据挖掘算法的改进,实现集中、分发数据的准确性与一致性。最后,将该算法应用到制造业的信息集成中,通过实证分析验证了该算法的有效性。

1 相关概念

1.1 概念格

概念格是对给定形式背景的三元组 $T=(O,D,R)$ 中事例集合 O,属性集合 D 之间形成的二元关系 R,R 与 O、D 之间偏序关系产生的一种诱导格。

建格的过程实际上是概念聚类的过程。因此,在概念格中,建格算法具有很重要的地位。对于同一批数据,所生成的格是唯一的,即不受数据或属性排列次序的影响,这也是概念格的优点之一。

1.2 ETL 工作流

本研究中 ETL (Extract-Transform-Load) 工作流的定义为 :通过基于工作流定义数据加工规则实现数据加工转换过程(包括抽取、清洗、转换、加载等过程)的业务规则,依据加工规则,实现工作流控制各类数据加工任务的执行在该工作流技术中要实现与概念格的结合,最重要的是要实现实时增量 ETL 过程。众多学者都对实时增量 ETL 技术进行了研究,提出了使用时间戳[7]、trigger[8]、日志[9] 等 3 种方法。从技术实现角度来看,以上 3 种技术相对成熟。实时增量 ETL 的关键问题在于,数据仓库数据是历史的、稳定的数据,而业务系统数据是时变数据,即某一条业务数据可能在一段时间内经历增、修、删以及可能的恢复等操作。因此实时抽取并加载,容易造成数据仓库汇总数据与实时业务数据之间的不一致问题,而这也是本研究要解决的关键问题。

* 收稿日期 :2012-09-17 修回日期 :2012-10-10 网络出版时间 :2013-03-16 13 :37
资助项目 :四川省自然科学基金(No. 11ZB152)
作者简介 :李志坚,男,助理研究员,硕士研究生,研究方向为计算机应用技术,E-mail :mysylizj@163.com
网络出版地址 :http ://www.cnki.net/kcms/detail/50.1165.N.20130316.1337.201302.92_021.html

2 基于概念格的数据挖掘算法改进

一般基于概念格的数据挖掘算法只是单纯的运用数据挖掘技术进行一系列发现。本研究将时间序列数据挖掘技术与 ETL 结合,以基于多元回归分析的时序趋势挖掘为例进行分析,即将样本数据获取、数据整理、数据预处理、复杂计算、偏相关分析、参数估计、变量选择、模型检验等均设计为 ETL 的动作,利用工作流技术,合理组织这些动作的执行次序及参数传递。

2.1 初始化工作

1)确定回归分析的因变量 Y,自变量 $X1,X2,\cdots,Xm$ 以及自变量个数 m;确定显著性水平;确定用于建立多元线性回归模型的样本数 n(当不知道 n 时,取全部原始数据);确定计算的精度(计算过程中,保留浮点数小数位的长度);确定用于回归分析的变量所在的数据源。

2)建立临时数据表 temp,如表 1 所示。

表 1　临时数据表 temp

序号	$X1$	$X2$	…	Xm	Y
1	$X11$	$X21$	…	$X1m$	$Y1$
2	$X12$	$X22$	…	$X2m$	$Y2$
…	…	…	…	…	…
n	$X1n$	$X2n$	…	Xnm	Yn

3)获取回归分析的原始数据,并将 $X1 \sim Xm$ 以及 Y 值导入临时数据表。

2.2 偏相关分析

根据 temp 表中的数据,依次计算因变量 Y 与各自变量 $Xi(1 \leqslant i \leqslant m)$ 的 $(m-1)$ 阶偏相关系数,放置到一维数组 $rp[m]$ 中,具体算法为定义数组 $m[]$,按大小序存放 $R_{ki}=1,2,\cdots,(i-1)(i+1),\cdots,m$ 中的序列 $ki=1,2,\cdots,(\mathrm{i}-1)(\mathrm{i}+1),\cdots,\mathrm{m}$,表示可用。计算 $R_{ki}=1,2,\cdots,(\mathrm{i}-1)(\mathrm{i}+1),\cdots,m$,直接调用 $R_p=P_R(0,i,m[])$。

1)函数 $P_R(k,i,m[])$。已知 Y 与 $Xi(i=1,2,\cdots,m)$)各样本值,用 $P_R(k,i,m[])$ 过程求 Y 与各 Xi 的 $m-1$ 阶偏相关系数。

2)函数 $R(k,i)$。因为递归会改变数组 m 中的标志位,所以建立新的数组 $flag[m.\,length-1]$,使 $flag[j]=m[j]$,j 从 0 到 $m.\,length-1$;$R(k,i)$ 求出 Xk 与 Xi 的简单相关系数 rki,并返回函数值。

2.3 变量选择(向前加入法)

1)根据显著性水平 a,查表得到可将变量引入回归方程的 F 引入的最小值 $Fmin$。

2)建立动态数组 Xj 和 Xs,分别记录已引入回归方程的自变量和尚未引入回归方程的自变量,Xj 初值为空,Xs 初值为 $\{X1,X2,\cdots,Xm\}$。

3)调用一元线性回归分析程序,对 Xs 中的每一个自变量与因变量 Y 的组合进行分析,得到回归方程的 t 检验的值 ti,放置到动态数组 T 中;根据显著性水平查表得到 ta 的值,放置到变量 ta 中。

4)根据动态数组 T 中的值,计算各自变量 Xi 中最大的 F 引入值及相应的 Xi,放置到变量 $Fmax$ 和变量 X 中。

5)若 $Fmax$ 值小于 $Fmin$ 或 $Fmax$ 值小于 ta,退出;否则记录变量 X 所在的回归方程,方程 t 检验值及 ta 的值,在 Xj 中添加 Xi(放置在变量 X 中),Xs 中删除 Xi。

6)分别将 Xs 中各自变量、与 Xj 中所有自变量及因变量 Y,根据 temp 表的数据进行分析,得到各回归方程;求各回归方程中新引入自变量 Xi(Xi 为数组 Xs 的元素)t 检验的值 ti,放置到动态数组 T 中;求各方程显著性检验值 F,放置到动态数组 F 中;及根据显著性水平查表得到的值 $Fa/2$,放置到变量 Fa 中。

7)根据动态数组 T 中的值,计算各新引入自变量 Xi 中最大的 F 引入值及相应的 Xi,放置到变量 $Fmax$ 和变量 X 中。

8)若 $Fmax$ 值小于 $Fmin$,退出;否则记录 F 引入值最大的自变量所在的回归方程及其 F 值(动态数组 F 中元素 $F[j]$,对应 xi 所在方程的 F 检验值)和 $Fa/2$ 值(放置在变量 Fa 中)。

9)根据 $F[j]<Fa$,退出;否则在 Xj 中添加 Xi(放置在变量 X 中),Xs 中删除 Xi。

10)循环,直到 Xs 为空。

2.4 趋势分析模型

首先,修改 temp 表,在字段“序号”后加入字段“$X0$”,其观测值均为“1”;其次,计算 XTX,放置到二维数组 $XtX[m+1][m+1]$ 中;计算 XTY,放置到一维数组 $XtY[m+1]$ 中;具体算法为

1)计算矩阵 XTX 的上三角部分(包括主对角线)和 XTY:

```
for(i=0;i<=m;i++)
 {
for(j=i;j<=m;j++)
 temp 表中 Xi 字段和 Xj 字段中对应的观测值相乘后累加;
 temp 表中 Xi 字段和 Y 字段中对应的观测值相乘后累加;
 }
```

其中,$tempi,k$ 表示 temp 表中 Xi 字段第 k 个观测值,$tempj,k$ 表示 temp 表中 Xj 字段第 k 个观测值。

2)利用矩阵 XTX 关于主对角线对称,填写下三角部分(不包括主对角线),算法为

```
for(i=0;i<=m;i++)
```

for($j=i;j<=m;j++$)

$XtX[j][i]=XtX[i][j]$。

3)依次计算回归方程中第 i 个($0\leqslant i\leqslant m$)自变量的统计量 t 并放置到一维数组 $T[i-1]$ 中。

4)查 t 分布表得到 $ta/2$,放置到变量 ta。

5)需要输出的结果:自变量 Xi 及其 t 检验的值 $T[i-1]$,ta。

6)根据 temp 表中的数据计算统计量 F,放置到变量 F 中。

7)查 F 分布表得到 $Fa/2$,放置到变量 Fa。

8)输出 F 和 Fa。

3 实证分析

基于上述的算法设计,在某汽车制造企业构建了基于 ETL 工作流的数据集成系统,并进行了实施和应用。系统由数据加工服务器、加工规则资料库以及数据加工规则开发设计工具构成。其中,数据加工服务器由 ETL Server 和 Meta Server 2 个服务组件组成。

资料库(Repository)保存 ETL 过程中的所有元数据,即数据仓库系统平台中的资料库(同时包含了其他相关产品的元数据),Repository 中的元数据同时可以采用 XML 文件加上嵌入式数据库的方式,可以不需要第三方的 DBMS 的支持,并且提供备份和恢复手段。ETL 元数据包括以下内容:1)据源定义:包括关系型数据库的连接参数、文本(XML)文件的存储位置 URI;2)目标数据库定义:目标数据库的位置定义;3)数据结构定义:包括数据源的数据结构描述和目标数据库数据结构的描述;4)ETL 逻辑规则:指抽取、清洗、转换、加载过程的业务规则,ETL 逻辑规则依赖于数据源和目标数据库的数据结构,与具体的数据源和目标数据库无关;5)ETL 运行规则:指 ETL 逻辑规则的实际运行计划,由 ETL 逻辑规则、数据源、目标数据库、执行周期、运行环境等元数据组成;6)ETL 执行日志:包括每一个 ETL 规则运行过程产生的所有信息。

实施过程如下包括以下 2 个步骤:1)初始化(图 1);2)偏相关预测及趋势分析(图 2)。

初始化过程主要包括以下步骤:对用户输入的数据 $x.$($x01$,$x02$,……,$x0m$)进行接收,并预测出计算点,将其放置到 $Y0$ 变量中,若预测时存在偏移量,则采用多元回归预测采用区间法执行。

偏相关预测及趋势分析的内容主要包括:根据 temp 表中的数据,依次计算因变量 Y 与各自变量 Xi($1\leqslant i\leqslant m$)的($m-1$)阶偏相关系数,放置到一维数组 $B[i]$ 中,运用数组 B 计算并得到多元回归方程。根据预测的结果对 temp 表中的数据进行修改,并计算出误差估计值,一直循环到可计算出修正自由度的可决系数,最后输出相应的结果则表示该步骤完成。

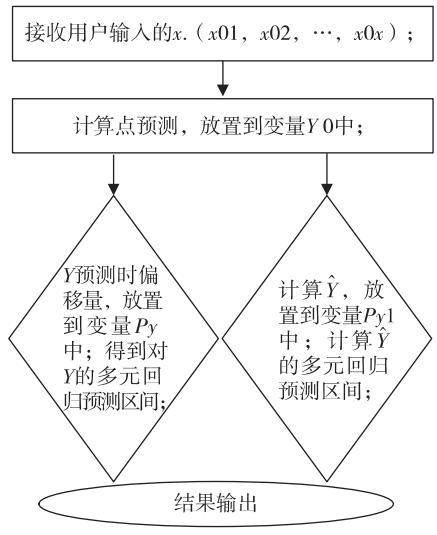

图1 初始化过程

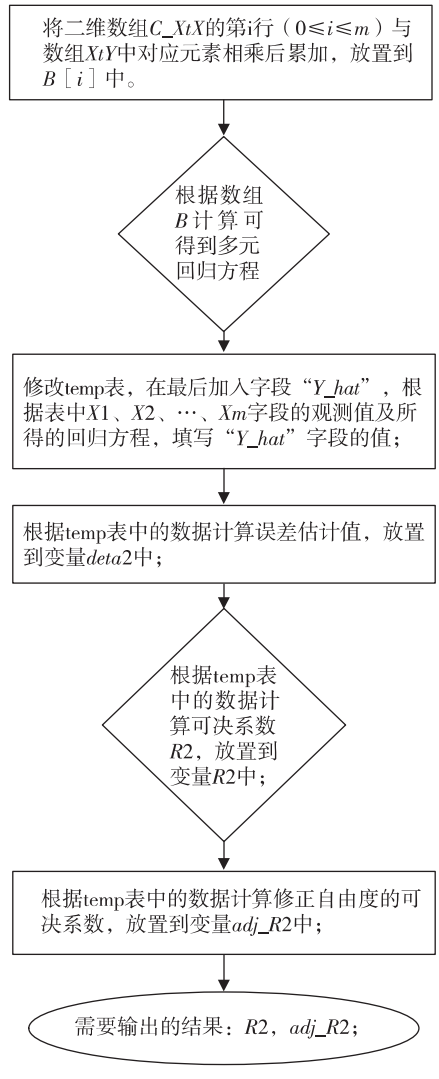

图2 偏相关预测及趋势分析过程

为了验证算法的有效性,对未实施系统前采用人工手段进行 ETL 动作处理与采用本系统进行 ETL 分解的效率进行了比较。

该实验在具有 2 GB 内存、P4 3.0 GHz 处理器的的 PC 机上进行,操作系统为 Windows XP。实验数据采用 2 组随机生成的样本数据。设定形式背景中对象平均拥有的属性数为 $\|D\|=40,60$,对象数为 $\|T\|=1\times10^9$,2×10^9,…,8×10^9,随机生成两组测试数据集。

为了更直观的反应该算法的精确度,将对象数均匀的分成 $n=2,4,8$ 个子数据集,图 3 是采用手工方法的处理精确度,图 4 是采用该算法后的处理精确度。从图 3 可以看出在数据越多的情况下,对象与精确度成正比,即验证了该算法对海量数据处理的有效性。

从图 3 和图 4 对比可以看出:随着 n 值的增大,手工处理的数据误差较大,变化的方式比较曲折,而采用

该算法处理的数据则误差很小,而且随着 n 增大折线呈现上升趋势,表示该算法精度越高。

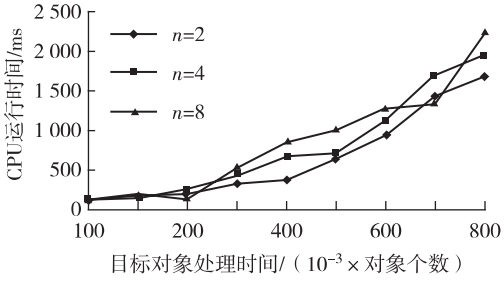

图3 采用手工方法处理的精确度

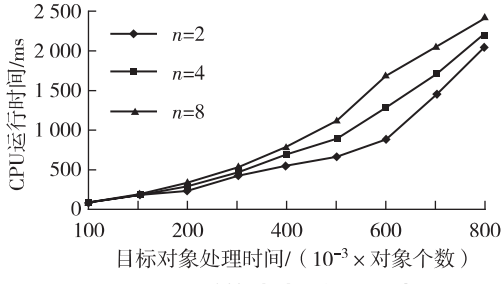

图4 采用该算法处理的精确度

4 结论

本研究将全局数据挖掘与局部数据挖掘相结合,基于局部信息的全局数据挖掘,通过局部信息的使用一方面获取算法效率的提高,另一方面通过局部信息的综合能够获得理解性更强的模型。该方法的创新之处为将挖掘过程分解为 ETL 动作,将时间序列数据挖掘技术与 ETL 结合,结合 ETL 处理工作流,实现并行分布式海量数据的时序挖掘。实验证明,该算法对增强数据加工能力具有一定的实用性。

参考文献:

[1] Yao Y Y. A comparative study of formal concept analysis and rough set theory in data analysis [C]//Proc of the rough sets and current trends in computing, LNCS. Berlin: Springer, 2007: 59-68.

[2] Zhang K, Hu Y F, Wang Y. An IRST-based algorithm for construction of concept lattices [J]. Journal of Computer Research and Development, 2009, 41(9): 1493-1499.

[3] 张文修,魏玲,祁建军. 概念格的属性约简理论与方法[J]. 信息科学,2008,6(35): 628-639.
Zhang W X, Wei L, Qi J J. Concept lattice attributes reduction theory and methods [J]. Information Science, 2008, 6(35): 628-639.

[4] Li H R, Zhang W X, Wang H. Classification and reductionof attributes in concept lattices [C]//Proc of IEEE int l conf on granular computing. Los Alamitos: IEEEComputer Society, 2006: 142-147.

[5] 薛峰. 基于概念格理论的粗集属性约简算法研究[D]. 合肥:合肥工业大学,2009.
Xue F. The Attribute reduction algorithms research of rough set based on concept lattice [D]. Hefei: Hefei University of Technology, 2009.

[6] Zhang W X, Wei L. Attribute reduction inconcept lattice based on discernibility matrix [C]//Proc of the 10th int l conf on rough sets, fuzzy sets, data mining and granular computing (RSFDGrC). LNCS 3642. Berlin: Springer, 2010: 157-165.

[7] Shao M W. The reduction for two kind of generalized concept lattice [C]//Proc of the 4th Int l conf on machine learning and cybernetics. Berlin: Springer, 2009: 2217-2222.

[8] 谭支鹏,冯丹,吴永英,等. 基于工作流的数据抽取转换加载[J]. 华中科技大学学报:自然科学版,2006,34(2): 61-63.
Tan Z P, Feng D, Wu Y Y, et al. Workflow-based data extraction conversion load [J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2006, 34(2): 61-63.

[9] 张旭峰,孙未未,汪卫,等. 增量 ETL 过程自动化产生方法的研究[J]. 计算机研究与发展,2010(6): 67-72.
Zhang X F, Sun W W, Wang W, et al. Incremental ETL process automation generation method research [J]. Journal of Computer Research and Development, 2010 (6): 67-72.

An Improved Concept Lattice-Based Data Mining Algorithm

LI Zhi-jian[1,2], MO Jian-lin[1,2]
(1. University of Electronic Science and Technology of China, Chengdu 610054;
2. Aba teachers college, Wenchuan Sichuan 623002, China)

Abstract: In order to solve the multidimensional data model and relational data model, query between the two-way data system, data cleansing, data conversion, distributed data accuracy and consistency control problem, this paper described the concept of grid-related, the global data mining combined with local data mining is proposed based on local information based on the concept of a global grid of data mining algorithm, and the mining process was divided into ETL action, combined with the ETL process workflow, using amounts of data distributed parallel sequence mining. Experiments show that the algorithm has a good effect on enhanced data processing capability.

Key words: data; concept lattice; data mining; integration; timing mining

(责任编辑 方 兴)