灰色线性回归模型在元规则挖掘中的应用研究*

曾庆飞,张忠林,刘丛林,梅玲霞
(兰州交通大学 电子与信息工程学院,甘肃 兰州 730070)

摘 要:提出了一种利用灰色线性回归组合模型挖掘关联规则元规则的方法,并通过实例分析证实了方法的有效性。

关键词:灰色线性回归;关联规则;元规则挖掘
中图分类号:TP311 **文献标识码**:A **文章编号**:1674-7720(2011)01-0010-03

Application of grey linear regression model in meta-association rules mining

Zeng Qingfei, Zhang Zhonglin, Liu Conglin, Mei Lingxia
(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: This paper puts forward one method of mining the meta-association rules for association rule which is based on grey linear regression model and through case analysis proves the validity of this method.

Key words: grey linear regression; association rules; meta-association rules mining

关联规则是数据挖掘领域应用非常广泛的挖掘方法,它主要用于发现事务数据集中项与项之间的关系,为决策者提供参考。基于经典关联规则的挖掘认为规则是永恒不变的,决策者只能利用这种静态规则信息进行分析和决策。实际上,规则并不一定永恒有效,例如:以某超市一年的销售数据库作为分析对象,有可能发现"顾客在购买香烟的同时也会购买礼品"这条规则,但通过分析数据库可知,支持这条规则的数据集大多集中在春节、圣诞节和国庆节前后,而在其他时间段规则支持度很小,并不具有全局指导作用。因此,利用基于静态宏观思想所挖掘出的规则进行决策存在一定的弊端。为了得到更加合理有效的决策信息,研究工作者提出了关联规则变化的挖掘。Abraham[1]首次提出了元挖掘的思想;荣冈等[2]提出了一种新的描述和评价关联规则的方法,从而为元规则定量预测分析提供了基础。本文将给出元规则形式化定义,并在参考文献[2]提出的支持度向量基础上利用灰色线性回归组合模型分析预测关联规则元规则。

*基金项目:甘肃省科技支撑计划项目(1011GKCA040);兰州市企业技术攻关项目(2009-1-4)

1 相关定义

1.1 元规则形式化定义

$P_1 \wedge P_2 \wedge \cdots \wedge P_t \Rightarrow Q_1 \wedge Q_2 \wedge \cdots \wedge Q_r$ 是元规则的规则模板,其中 $P_i(i=1,2,\cdots,l)$ 和 $Q_j(j=1,2,\cdots,r)$ 是例示谓词或谓词变量。元规则是对数据间依赖关系的关联规则的形式模式,它使用户可以说明其感兴趣的规则的语法形式,规则的形式可以做为约束,指导知识的发现,帮助提高挖掘过程的性能。

1.2 支持度向量

任务相关的事务数据集 D 是在时间段 t 内收集到的,时间段 t 可以分成不相交的、长度为 n 的时间序列 $\{t_1,t_2,\cdots\cdots,t_n\}$,因此,数据集 D 可以相应地划分为 n 个子数据集 $\{D_1,D_2,\cdots\cdots,D_n\}$。关联规则 $A \Rightarrow B$(或者项集 $A \cup B$)支持度向量,具有如下表示形式:

$$SV=[S_{(A\cup B)1},S_{(A\cup B)2},\cdots,S_{(A\cup B)n}]\text{st.}S_{(A\cup B)i}=f_{(A\cup B)i}/|D_i| \quad (i\in\{1,2,\cdots,n\}) \tag{1}$$

其中 $f_{(A\cup B)i}$ 为项集 $A\cup B$ 在数据子集($D_i\in\{1,2,\cdots,n\}$)中出现的频数,$|D_i|$ 为 D_i 中的事务数。

设规则 $A\Rightarrow B$ 的支持度为 S,则:

$$S=S_{(A\cup B)}=f_{(A\cup B)}/M=\sum_{i=1}^{n}f_{(A\cup B)i}/M \tag{2}$$

其中 M 为 D 中的事务数。

2 灰色线性回归组合模型建模方法

元规则挖掘是针对单个规则的信息进行分析和预测，对每一条相同的规则根据不同的时间粒度划分数据库可以建立不同的数据序列。针对超市销售数据库、电信客户数据库等，以小时间粒度划分数据库进行分析的意义不是很大，一般按照年、月、周进行数据划分，因此数据建模序列通常并不是十分复杂，适合用灰色理论进行研究。目前提出建立元规则的方法主要有基于概率统计的方法[3]和基于模糊决策树的方法[4]。基于概率的方法主要采用主成份分析、回归分析等对规则的支持度进行曲线拟合，这在处理不确定数据上效果欠佳；而基于模糊决策树的方法由于需要较多的专家信息，明显无法满足要求。对于具备线性和指数趋势的小样本序列，灰色线性回归组合模型是一种很好的数据预测模型，其建模过程如下[6]：

(1)设序列 $X^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n))$，对 $X^{(0)}$ 进行一次累加生成处理，得到生成序列 $X^{(1)}=(x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n))$，其中 $x^{(1)}(i)=\sum_{t=1}^{i}x^{(0)}(t)\ \ (i=1,2,\cdots,n)$。

(2)按照传统 GM(1,1)建模过程可得：

$$\hat{x}^{(1)}(t+1)=(x^{(0)}-b/a)\exp(-at)+b/a$$

其中 a 为发展系数，b 为灰色作用变量。其形式可记为 $\hat{x}^{(1)}(t+1)=C_1\exp(\nu t)+C_2$，用线性回归方程 $Y=aX+b$ 及指数方程 $Y=a\times\exp(X)$ 的和来拟合累加生成序列 $x^{(1)}(t)$，因此可将生成的序列写成：$\hat{x}^{(1)}(t)=C_1\exp(\nu t)+C_2t+C_3$，其中参数 ν 及 C_1、C_2、C_3 待定。

(3)为了确定以上参数，设参数序列 $z(t)=\hat{x}^{(0)}(t+1)-\hat{x}^{(1)}(t)=C_1\exp(\nu t)[\exp(\nu)-1]+C_2,(t=1,2,\cdots\cdots,n-1)$。设 $y_m(t)=z(t+m)-z(t)$，$y_m(t+1)=C_1\exp[\nu(t+1)][\exp(\nu_m)-1][\exp(\nu)-1]$，两式之比为：

$$y_m(t+1)/y_m(t)=\exp(\nu) \tag{3}$$

(4)将步骤(3)中的 $\hat{x}^{(1)}$ 换成 $x^{(1)}$，得到 ν 的近似解 $\tilde{V}$，取不同的 m 可得到不同的 $\tilde{V}$，以它们的平均值作为 ν 的估计值 $\hat{V}$，计算公式为：

$$\hat{V}=\frac{\sum_{m=1}^{n-3}\sum_{t=1}^{n-2-m}\tilde{\nu}_m(t)}{(n-2)(n-3)/2} \tag{4}$$

(5)设 $l(t)=\exp(\hat{V}t)$，则 $\hat{x}^{(1)}(t)=C_1l(t)+C_2t+C_3$，令 $x^{(1)}=[x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n)]^{\mathrm{T}}$，$C=[C_1,C_2,C_3]^{\mathrm{T}}$，$A=\begin{bmatrix}l(1)&1&1\\l(2)&2&1\\\cdots&\cdots&\cdots\\l(n)&n&1\end{bmatrix}$，从而

$$C=(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}X^{(1)} \tag{5}$$

对生成序列预测值 $\hat{x}^{(1)}(t)$，按照公式 $\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)$ 累积还原得到原始序列预测值。

3 实例分析

本文以某通信公司 2008 年的客户数据库的业务记录为原始基础数据，按照月份将数据集划分为 12 个子数据集，并利用参考文献[2]提出的关联规则挖掘算法挖掘得到频繁项目集。分析由频繁 2 项集生成的一条关联规则“固定电话业务=>163 拨号业务”(即客户在办理固定电话业务的前提下同时办理 163 拨号业务)的规则变化情况。该规则每月的支持度计数构成规则支持度向量 $SV=$[72,85,90,103,117,126,155,168,193,224,265,308]，选取规则前十个月的支持度数据作为建模原始数据，将 11 月和 12 月的数据作为模型有效性检验数据。下面分别用灰色线性回归组合模型[5]和线性回归模型[6]进行预测分析。

3.1 线性回归模型

(1)当对事务数据库引入时间因素后，规则支持度计数和时间就存在了密切关系，设规则支持度计数为因变量 Y_i，月份为自变量 X_i，根据前十个月统计资料做散点图如图 1 所示。

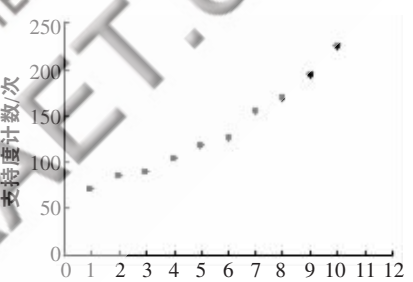

图 1 规则支持度计数随月份变化的散点图

(2)设最佳拟合回归直线方程为 $\hat{Y}=a+bX$，由一元线性回归直线方程参数计算公式：$b=\dfrac{\sum_{i=1}^{n}X_iY_i-\bar{x}\sum_{i=1}^{n}Y_i}{\sum_{i=1}^{n}X_i^2-\bar{x}\sum_{i=1}^{n}X_i}$ 及 $a=\bar{y}-b\bar{x}$(其中 $\bar{x}=\sum_{i=1}^{n}x_i/n$，$\bar{y}=\sum_{i=1}^{n}Y_i/n$)，得 $a=44$，$b=16.236$，回归方程为 $\hat{Y}=44+16.236X$。

3.2 灰色线性回归组合模型

(1)原始序列为 $x^{(0)}=\{72,85,90,103,117,126,155,168,193,224\}$，由公式 $x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i)$ 得累加生成序列 $x^{(1)}=\{72,157,247,350,476,593,748,916,1\ 109,1\ 333\}$。

(2)应用公式(3)、(4)可得到 ν 的估计值 $\hat{V}=0.184\ 3$。

(3)根据公式(5)可得待估参数 $C_1=170.103\ 4$，$C_2=43.621\ 1$，$C_3=-177.013\ 9$，于是一次累加生成序列的组合模型为：$\hat{x}^{(1)}(t)=170.103\ 4\exp(0.184\ 3t)+43.621\ 1t-177.013\ 9$。

(4)将得到的累加预测值$\hat{x}^{(1)}(t)$按照公式$\hat{x}^{(0)}(t+1)=\hat{x}^{(1)}(t+1)-\hat{x}^{(1)}(t)$累积还原即可得到原始值的预测结果。

3.3 模型拟合及预测结果比较

依据上面所述线性回归模型和灰色线性回归模型求解步骤,分别计算两种模型的预测值,如表1所示(预测值均取整数)。利用相对误差法检验两种模型均满足精度要求,可以用于进一步预测。由表1可知线性回归模型拟合结果平均相对误差和预测结果平均相对误差分别为6.96%、19.16%,灰色线性回归模型拟合和预测的相对误差分别为1.37%、1.24%,灰色线性回归拟合和预测精度均明显优于线性回归模型。

表1 模型误差检验表

月份	实际值	线性回归模型		灰色线性回归模型	
		预测值	相对误差/%	预测值	相对误差/%
1	72	60	16.67	71	1.39
2	85	76	10.59	85	0.00
3	90	93	3.33	93	3.33
4	103	109	5.82	103	0.00
5	117	125	6.84	116	0.85
6	126	141	11.90	131	3.97
7	155	157	1.29	151	2.58
8	168	174	3.57	170	0.60
9	193	190	1.55	194	0.52
10	224	206	8.04	225	0.45
11	265	223	15.85	261	1.51
12	308	239	22.40	305	0.97

图2通过图示进一步直观地对两种预测模型进行比较可知,灰色线性回归模型的预测值与实际值相比,波动范围较小,图形更吻合,预测精度更好。灰色线性回归模型在动态关联规则元规则挖掘上具有良好的有效性,可以应用于实际分析中。由组合模型预测结果可知,此规则的有效性随着时间推移在不断地增强,在后续的时间中应该有很好的适用性,决策者可以对办理固定电话业务的客户推荐163拨号业务。

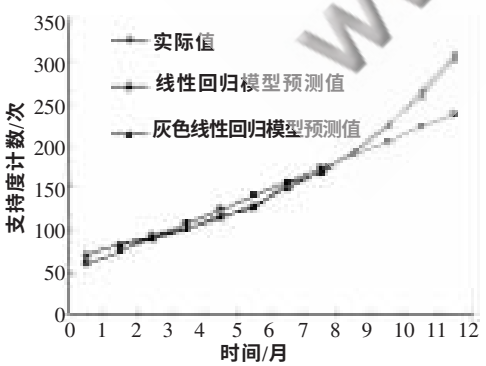

图2 两种预测模型的比较

本文提出了一种灰色线性回归组合模型的关联规则元规则挖掘方法,弥补了静态关联规则无法提供规则自身变化的不足,并能够对关联规则元规则变化的假定和判断基于时序数据的定量分析和研究。通过挖掘通信公司客户数据库业务数据,并利用不同的预测模型对规则支持度预测结果比较分析表明:灰色线性回归组合模型对具有线性和指数趋势的规则时间序列的拟合及预测精度均优于线性回归模型,从而可以更加准确地反映规则的变化趋势,判断规则的有效性,使决策者正确把握规则在实际中的应用前景。

参考文献

[1] ABRAHAM T, RODDICK J F. Incremental meta-mining from large temporal data sets. Advances in Database Technologies, Proceedings of the 1st International Workshop on DataWarehousing and Data Mining (DWDM'98), 1999:41-54.

[2] 荣冈,刘进锋,顾海杰.数据库中动态关联规则的挖掘[J].控制理论与应用,2007,24(1):127-131.

[3] Liu Bing, Ma Yiming, Lee R. Analyzing the interestingness of association rules from the temporal dimension [J]. IEEE International Conference on Data Mining (ICDM-2001), Silicon Valley, CA, 2001.

[4] Wai-Ho Au, Keith C. C. Chan. Mining changes in association rules: a fuzzy approach [J]. Fuzzy sets and systems, 2005,149(1): 87-104.

[5] 刘思峰,党耀国,方志耕,等.灰色系统理论及其应用[M].北京:科学出版社,2004:125-138.

[6] 王松桂.线性回归与方差分析[M].北京:高等教育出版社,1999.

(收稿日期:2010-09-15)

作者简介:

曾庆飞,男,1985年生,硕士研究生,主要研究方向:数据挖掘。

张忠林,男,1965年生,教授,博士,主要研究方向:数据挖掘、软件工程。