基于小波变换的动态关联规则元规则GM(1,1)挖掘

张忠林 许 凡
(兰州交通大学电子与信息工程学院 兰州 730070)

摘 要 提出了一种把小波变换应用到动态关联规则元规则挖掘中并提高规则预测精度的方法。该方法首先利用小波变换技术对挖掘出的动态关联规则元规则支持度计数进行变换,然后通过小波变换的多分辨率特点提取出近似部分和细节部分,并利用这两部分别进行单支重构,随后利用GM(1,1)对重构的两部分进行预测,从而得到最后的预测结果,最后通过实验证明了该方法具有较高的预测精度。

关键词 小波变换,动态关联规则,元规则,单支重构,GM(1,1)
中图法分类号 TP273 **文献标识码** A

Meta-association Rule Mining for Dynamic Association Rule Used GM(1,1) Based on Wavelet Transform

ZHANG Zhong-lin XU Fan
(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract This paper put forward a method for applying wavelet transform to meta-rule mining in dynamic association rules to improve the forecast accuracy of the rules. First, it uses Daubechies wavelet to transpose the support for the technology of the meta-rule in dynamic association rules. And then it extracts the approximate part and detail part according to the multi-resolution characteristics of wavelet transform. Finally by single reconstructing the two parts the GM(1,1) can be used to forecast the reconstructed two parts to get the final prediction results. The last experiment gives higher prediction accuracy.

Keywords Wavelet transform, Dynamic association rules, Meta-association rule, Single reconstructed, GM(1,1)

1 概述

在数据挖掘中关联规则是最近几年研究得比较多的话题,1993年首先由Agrawal R提出关联规则的概念,但是所挖掘出来的数据没有考虑随着时间变化而变化的,因为关联规则在时间上是有规则变化的,例如周期性、趋势性等,基于此原因,GANTI V和GEHRKE J在2001年首次提出了关联规则的时间特性。而国内文献[1]提出了用支持度和置信度两个特性来衡量关联规则在时间上变化的动态性,紧接着在文献[2]中给出了动态关联规则的定义,利用动态关联规则可以分析模式自身的变化过程和预测规则未来的发展趋势,而文献[3]改进了动态关联规则定义的缺陷,并且给出了两种改进的动态关联规则挖掘算法,随后文献[4]提出了关于动态关联规则及其元规则基于时间序列趋势变化的挖掘。小波变换由于具有多分辨率,即多尺度的特点,可以由粗及精地逐步观察信号的变化,同时小波分析具有优越的时频域分析能力,因而成为非平稳信号处理的强有力的工具。小波变换已被应用于股指预测、网络流量预测等各个领域[5-7]中,取得了很好的效果,文献[8]将小波变换应用到时间序列挖掘中,通过小波分解一层一层分解到不同的频率通道上,分解后的时间序列在频率成分上比原始信号单一,并且小波分解对时间序列做了平滑,因此利用小波变换的多分辨率的特点对挖掘得到的大量数据支持度计数向量进行变换,可以观测近似逼近部分的变化特征。而灰色系统理论[9,10]是建立系统运行趋势模型的有效方法,适用于动态预测,且只需少量已知信息就可建立预测模型[11],因此把动态关联规则和小波变换、灰色预测模型结合起来能提高预测的精准度[12]。

2 动态关联规则

关联规则是数据挖掘领域应用非常广泛的一种挖掘方法,但关联规则在时间上并不是不变的,关联规则是随着时间的变化而变化的,其支持度和置信度在时间上构成的向量可以反映关联规则在时间上的变化趋势等。

设$I=\{i_1,i_2,\cdots,i_n\}$是项集合,任务相关的事务数据集D是在时间段t内收集到的,$t=\{t_1,t_2,\cdots,t_n\}$可分为相等不相交的长度为n的时间序列,因此整个数据集D可以根据时间段t的划分为n个数据子集$D=\{D_1,D_2,\cdots,D_n\}$,而其数据子集$D_i(i\in\{1,2,\cdots,n\})$的数据是在$t_i(i\in\{1,2,\cdots,n\})$时间段内收集的项集$T$,并且满足$T\subseteq I$。若$A$和$B$为项集,$A\subset I$,$B\subset I$,并且$A\cap B=\emptyset$,则有如下动态关联规则的相关定义。

到稿日期:2012-07-12 返修日期:2012-10-16 本文受国家自然科学基金项目(61163010),甘肃省科技支撑计划项目(1011GKCA040)资助。
张忠林(1965-),男,博士,教授,CCF高级会员,主要研究方向为智能信息处理、软件工程,E-mail:zhangzl@mail.lzjtu.cn;许 凡(1987-),男,硕士生,主要研究方向为数据挖掘。

定义 1 支持度向量(SV)是动态关联规则 $A\Rightarrow B$(或者项集 $A\cup B$)的向量。

支持度向量具有如下的表示形式

$$SV=[s_{(A\cup B)_1},s_{(A\cup B)_2},\cdots,s_{(A\cup B)_n}]$$

$$S_{(A\cup B)_i}=f_{(A\cup B)_i}/|D_i|(i\in\{1,2,\cdots,n\})$$

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

定义 2 设项集 $A\cup B$ 的支持度为 s,则

$$s=s_{(A\cup B)}=f_{(A\cup B)}/M=\sum_{i=1}^{n}f_{(A\cup B)_i}/M$$

式中,M 是 D 中的事务数。有时,项集出现的频数表示支持度更为合适,这样项集的支持度向量为 $SV=[f_1,f_2,\cdots,f_n]$。

而元规则是形如以下的规则模式

$$A_1\wedge A_2\wedge\cdots\wedge A_l\Rightarrow B_1\wedge B_2\wedge\cdots\wedge B_r$$

式中,$A_i(i=1,2,\cdots,l)$ 和 $B_j(j=1,2,\cdots,r)$ 是示例谓词或是谓词变量。

定义 3 在事务数据 $D_i=\{D_1,D_2,\cdots,D_n\}(i=1,2,\cdots,n)$ 上,规则 $A\Rightarrow B$ 的支持度向量定义为:$SV=[s_{(A\cup B)_1},s_{(A\cup B)_2},\cdots,s_{(A\cup B)_n}]$ 只有当 $S_{(A\cup B)_i}\geqslant$min_sup,其中 min_sup 为最小支持度。则动态关联规则在数据集 D 上的支持度元规则为 $A\Rightarrow B:SV$。

3 小波变换

3.1 小波变换层次分解

小波变换分解的最终目的是力求构造一个在频率上高度逼近 $L^2(R)$ 空间的正交小波基,小波变换的多分辨率对近似部分和细节部分进行进一步的分解,因此使频率的分辨率变得越来越高。从函数空间的分解上引入多分辨率。

(1)离散逼近:将平方可积函数 $f(t)\in L^2(R)$ 看成是某一逐级的极限情况,每一级的逼近都是用某一低通平滑函数 $\Phi(t)$ 对 $f(t)$ 做平滑的结果,在逐级逼近时,平滑函数 $\Phi(t)$ 也做逐级伸缩,这就是多分辨率,即用不同分辨率来逼近带分析的函数 $f(t)$,在空间上描述为对空间做逐级二分解产生一组逐级包含的子空间,如下所示

$$V_0=V_1\oplus W_1,V_0=V_2\oplus W_2,\cdots,V_j=V_j\oplus W_j$$

式中,j 是从$-\infty$到$+\infty$的整数,j 的数值越小代表空间越大,而 V_j 与 W_j 是正交补空间,当 $j\to-\infty$ 时,$v_j\to R^2$,包含为整个平方可积函数的实变函数空间。即 $\bigcup_{j\in Z}V_j=L^2(R)$,当 $j\to+\infty$,$v_j\to\langle 0\rangle$,即空间最终分解到空集为止,在逐级包含的情况下 $\bigcap_{j\in Z}V_j=\langle 0\rangle$,并且 V_j 与 W_j 是正交补空间,同时 $V_0\supset V_1\supset V_2\supset V_3\supset\cdots\supset V_j$ 和 $V_0=V_1\oplus W_1=V_2\oplus W_2\oplus W_1=\cdots=V_j\oplus W_j\oplus\cdots W_2\oplus W_1$。

这种函数空间具有以下特性:

①位移不变性:函数的时移不改变其所属空间,即

如果 $f(t)\in V_j$ 则 $f(t-k)\in V_j$

②二尺度伸缩性:即 $f(t)\in V_j$,则 $f(\frac{t}{2})\in V_{j+1}$ 和 $f(2t)\in V_{j-1}$,设 V_0 中有低通平滑函数 $\Phi(t)$,它的整数移位集合 $\{\Phi(t-k)\}_{k\in Z}$ 是 V_0 中的正交归一基,则可称 $\Phi(t)$ 为尺度函数,所以有:$\langle\Phi(t-k),\Phi(t-k')\rangle=\Phi(k-k')$,其中 $\Phi(t-k)=\Phi_{0k}(t)$,$\Phi_{0k}(t)$ 为当 $j=0$ 时的 $\Phi_{jk}(t)=2^{\frac{-j}{2}}\Phi(2^{-j}t-k)$,$V_0$ 中的任意函数 $f(t)$ 可以表示为 $\{\Phi(t-k)\}_{k\in Z}$ 的线性组合,设 $p_0f(t)$ 代表 $f(t)$ 在 V_0 上的投影,则有 $p_0f(t)=\sum_k x_k^{(0)}\Phi_{0k}(t)$,其中 $x_k^{(0)}$ 是线性组合的权重,它的求法如下所示

$$x_k^{(0)}=\langle p_0f(t),\Phi_{0k}(t)\rangle=\langle f(t),\Phi_{0k}(t)\rangle$$

则称 $p_0f(t)$ 为 $f(t)$ 在 V_0 处的平滑逼近,也就是 $f(t)$ 在分辨率 $j=0$ 的概貌,$x_k^{(0)}$ 称为 $f(t)$ 在分辨率 $j=0$ 的离散逼近。

(2) 细节差异:假设在子空间 W_0 中能找到一个带通函数 $\Psi(t)$,其整数移位集合 $\{\Psi(t-k)\}_{k\in Z}$ 是 W_0 中的正交归一基,同样根据二尺度的伸缩性可得 $\Psi(t)\in W_0$,则尺度为 $\frac{t}{2}$ 属于空间 W,即 $\Psi(\frac{t}{2})\in W_1$,并且 $\{\Psi_{1k}=\frac{1}{\sqrt{2}}\Psi(\frac{t}{2}-k)\}$,必然构成 W_1 的一组正交归一基,$\langle\Psi(t-k),\Psi(t-k')\rangle=\delta(k-k')$,$W_1$ 中的任意函数 $f(t)$ 可以表示为 $\{\Psi(t-k)\}_{k\in Z}$ 的线性组合,设 $D_0f(t)$ 代表 $f(t)$ 在 W_1 上的投影,则有 $D_1f(t)=\sum_k d_k^{(1)}\Psi_{1k}(t)$,其中 $d_k^{(1)}$ 是线性组合的权重,它的求法如下所示

$$d_k^{(1)}=\langle D_1f(t),\Psi_{1k}(t)\rangle$$

并且 $p_0f(t)=p_1f(t)+D_1f(t)$,即 $D_1f(t)$ 是 V_0 与 V_1 二级的相邻平滑逼近之差,反映了这二级的逼近间的误差,因此把 $d_k^{(1)}$ 称为在分辨率 $j=1$ 的离散细节,而 $\Psi(t)$ 就是具有带通性质的小波函数,同时可以把 $d_k^{(j)}$ 称为小波变换 $WT_f(j,k)$。

(3)由于 $p_0f(t)=\sum_n x_n^{(0)}\Phi_{0n}(t)$,且因为 $D_1f(t)$ 与 $\Phi_{1k}(t)$ 正交,则 $\langle D_1f(t),\Phi_{1k}(t)\rangle=0$,即 $x_k^{(1)}=\langle\sum_n x_n^{(0)}\Phi_{0n}(t),\Phi_{1n}(t)\rangle=\sum_n\langle\Phi_{0n}(t),\Phi_{1k}(t)\rangle x_n^{(0)}$ 而 $\langle\Phi_{0n}(t),\Phi_{1k}(t)\rangle=\frac{1}{\sqrt{2}}\int\Phi(\frac{t}{2}-k)\Phi(t-n)\mathrm{d}t$,因此 $x_k^{(1)}=\sum_n h_{0(n-2k)}x_n^{(0)}$,同样可推导出 $d_k^{(1)}=\sum_n h_{1(n-2k)}x_n^{(0)}$,从而可以导出二层分解结构图如图 1 所示。

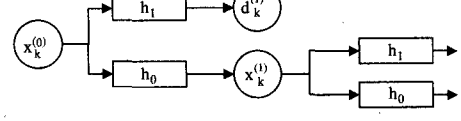

图 1 二层分解图

3.2 单支重构

单支重构是指不对近似部分和细节部分同时进行重构,而是对它们分别进行重构,即在对近似部分进行重构时将细节部分置零。重构的公式如下所示

$$x_n^{(j-1)}=\sum_k h_{0(n-2k)}x_k^{(j)}+\sum_k h_{1(n-2k)}d_k^{(j)}=\sum_k g_{0(n-2k)}x_k^{(j)}+\sum_k g_{1(n-2k)}x_k^{(j)}$$

上式为相邻二级的反演关系,其中 $x_k^{(j)}$ 和 $d_k^{(j)}$ 是第 j 级的离散平滑和离散细节逼近,而 $x_n^{(j-1)}$ 是由这两个离散逼近重构得到的,同时 $g_{0(n-2k)}$、$g_{1(n-2k)}$ 和 $h_{0(n-2k)}$、$h_{1(n-2k)}$ 一样。因此得到的重构过程(以二层为例)如图 2 所示。

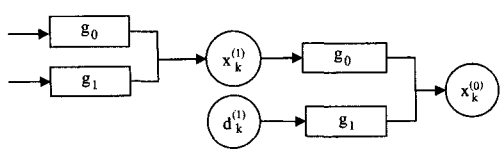

图 2 二层重构图

设 $Y=\{x_1,x_2,\cdots,x_n\}$ 即 $Y=SV$，是动态关联规则及其元规则的支持度计数向量在时间上变化的信号，对信号进行小波变换后分解为：离散逼近部分，即为近似部分 $A=\{A_1,A_2,\cdots,A_i,\cdots,A_n\}$；离散细节逼近部分，即为细节部分 $D=\{D_1,D_2,\cdots,D_i,\cdots,D_n\}$，其中 $i=\{1,2,\cdots,n\}$，i 代表分解的层次，n 为正整数，近似部分 A 代表了整个时间序列信号的整体特征和趋势变化，近似部分的信号也叫低频信号；而细节部分信号代表着整个时间序列信号随机噪声信息，也称为高频信号。近似部分 A 中的 $A_i=\{a_{i,1},a_{i,2},\cdots,a_{i,n}\}$ 代表第 i 层的近似信号，而细节部分 D 中的 $D_i=\{d_{i,1},d_{i,2},\cdots,d_{i,n}\}$ 代表第 i 层的细节信号。

4 *GM*(1,1)模型

4.1 *GM*(1,1)模型建立

(1)设时间序列 $X^{(0)}$ 有 n 个观察值，$X^{(1)}=\{x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n)\}$，为了使其成为有规律的时间序列数据，对其作一次累加生成运算，即令

$$x^{(1)}(t)=\sum_{n=1}^{t}x^{(0)}(n) \tag{1}$$

从而得到新的生成数列 $X^{(1)}$，即 $X^{(1)}=\{x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n)\}$。

(2)一般将新数列近似地服从指数规律，则生成的离散形式的微分方程具体的形式为

$$\frac{dx}{dt}+ax=u \tag{2}$$

式中，a,u 是发展系数和灰作用量，由最小二乘估计方法得

$$a=\begin{pmatrix}a\\u\end{pmatrix}=(B^TB)^{-1}B^TY \tag{3}$$

(3)构造数据矩阵。式(3)中 Y 为列向量，$Y=[x^{(0)}(2),x^{(0)}(3),\cdots\cdots,x^{(0)}(n)]^T$，$B$ 为数据矩阵，

$$B=\begin{pmatrix}-\frac{1}{2}[x^{(1)}(1)+x^{(1)}(2)] & 1\\ -\frac{1}{2}[x^{(1)}(2)+x^{(1)}(3)] & 1\\ \vdots & \\ -\frac{1}{2}[x^{(1)}(n-1)+x^{(1)}(n)] & 1\end{pmatrix} \tag{4}$$

(4)求出预测模型

$$X^{(1)}(t+1)=\left[X^{(0)}(1)-\frac{u}{a}\right]e^{-au}+\frac{u}{a} \tag{5}$$

4.2 *GM*(1,1)模型的检验

灰色预测即是通过对原始数据的累加处理和灰色模型的建立进而发现系统发展规律，从而对系统的未来趋势做出的预测。选择的预测模型需经过多重检验后才能判定其是否合理有效，只有通过检验的模型才能用于实际的预测。本文采用残差大小检验，即对实际值和实际值的残差进行逐点检验。

设模拟值的残差序列为 $e^{(0)}(t)$，则

$$e^{(0)}(t)=x^{(0)}(t)-\hat{x}^{(0)}(t) \tag{6}$$

令 $\varepsilon(t)$ 为残差相对值，即残差百分比为

$$\varepsilon(t)=\left[\frac{x^{(0)}(t)-\hat{x}^{(0)}(t)}{x^{(0)}(t)}\right]\% \tag{7}$$

模型建立后，同时需进行模型适用范围的检验。*GM*(1,1)模型适用范围与发展系数 a 相关[13,14]，如表 1 所列。

表 1　模型适用范围表

−a	模型适用范围
−a< 0.3	可用于中短期预测
0.3<−a< 0.5	可用于短期预测
0.5<−a< 0.8	可用于短期预测，应十分谨慎
−a>1	不宜采用

5 实验处理过程

输入：min_sup//最小支持度，数据集 D

数据集 D 录入过程：

(a)首先设计一个 Transaction 事务类，该类的对象包含的属性为 Id 整型数据；项集 item-set 字符串(形如“I1 I2 I8 I14”)，时间 date 类型 DATE 。

(b)main()//主程序

{InsertoSql();//调用插入事务数据库的函数}

{void InsertoSql();

{For(int i=0;i<50000;i++)//50000 循环

{1. 利用 java 的随机函数类 Random、简单日期格式类 SimpleDateFormat、日历类 Calendar 随机生成在一定时间范围内的 5000 条事物对象。

2. Transaction ts = new Transaction();//每次循环生成一个 Transaction 对象。

3. 对生成的 ts 对象的属性值分别赋值 //例如 item-set 为“I1 I5 I8”，时间 date 为 yyyy-MM-dd，代表年月日，这里的唯一标识在数据库里面为自动增长，自动生成。

4. Save(ts)//保存事务即插入数据库中，这里利用的是 hibernate 框架连接的 Mysql 数据库。

}}}

输出：频繁频繁项集 L 及其对应的支持度向量计数 SV

首先利用 FP-growth 算法找到所有的频繁项集 L

(1)for each D_i do{

//扫描数据库

(2)for each transaction t∈D_i do{

//扫描数据库

(3)L_{temp}=subset(L,t)//得到 t 所包含的是频繁项集的子集

(4)for each frequent item-set $l_j\in L_{temp}$ do

{$f_{(AUB)_i}$++;}}}

(5)由(4)得到的 $f_{(AUB)_i}$，然后再利用前面第 2 节提到的公式 $S_{(AUB)_i}=f_{(AUB)_i}/|D_i|$(i∈{1,2,…,n})来进行运算，得到支持度计数向量 SV，若 $Y=\{x_1,x_2,\cdots,x_n\}$ 即 Y=SV 是需要进行预测的时间序列信号，先对其进行小波分解。

(6)利用小波变换对得到的近似部分 A_i 和细节部分 D_i 分别进行单支重构，从而得到重构后的序列 $Y'=D_1'+D_2'+\cdots+D_i'+A_i'$，即，其中 A_i' 是 A_i 重构后的信号，同样 D_i' 是 D_i 重构后的信号。

(7)分别对 A_i' 和 D_i' 建立灰色 GM(1,1)预测模型，然后再进行误差分析检验。

(8)利用得到的 Y'(P)与 Y 进行计算，即 $\delta=\sum_{i=1}^{n}X_i|X_i-X_i'(P)|$ 得到拟合值与实际值的差额，以及差额所占的相对误差百分比。

6 数据分析

为了直观地说明上述方法在动态关联规则的元规则挖掘中预测支持度计数值的具体过程，下面将针对一实例进行说明。设 $I=\{i_1,i_2,\cdots,i_{15}\}$ 是项集合，对由这 15 个项随机生成的 50000 条数据构成的事务数据集 D 在总的时间长度上进

行2000个等时间段的分割，设定最小支持度 min_sup 为4%，即含义为在50000条事务数据集 D 中，挖掘出最少含有2000条的形如 $A\Rightarrow B$ 的规则，首先利用第5节中的算法过程(1)－(5)对数据库中的数据进行挖掘，并计算其支持度计数向量 SV，其中得到大于最小支持度 min_sup 的规则有111条，以其中的规则 $i_{10}\Rightarrow i_7$ 为例，在数据库所有的事务中总共有8105条包含此规则，在实际生活中，若该数据为某商场的销售数据，则其含义为消费者购买 i_{10} 产品之后购买 i_7 产品的总支持度计数值为8105/50000，规则 $i_{10}\Rightarrow i_7$ 的2000个支持度计数向量 SV 的值随着时间变化的结果如图3所示。

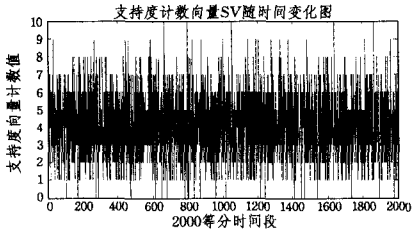

图3　支持度计数向量 SV 随时间变化图

然后对得到的规则 $i_{10}\Rightarrow i_7$ 的2000个支持度计数向量 SV 进行小波变换，用 db2 小波进行变换并进行4层分解，然后分别对近似部分 A 和细节部分 D 进行单支重构，则近似部分 $A4$ 单支重构的结果图如图4所示。

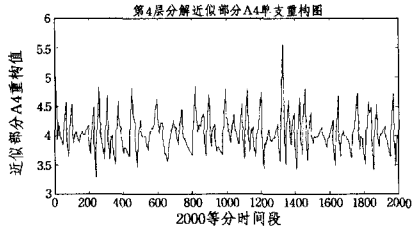

图4　近似部分 A4 重构结果图

由于小波分解后的近似部分代表原始信号的相似逼近部分，因此相对图3来说，从图4可以看出，规则在时间上的变化特征，而各层的细节部分 D 单支重构后的结果如图5所示。

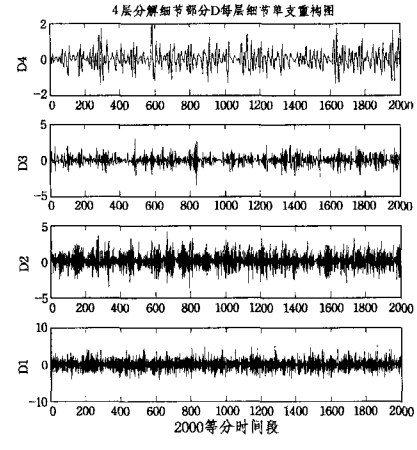

图5　细节部分重构图

利用小波变换单支重构方法得到重构后的值为 $Y'=D_1'+D_2'+D_3'+D_4'+A_4'$，其中 A_4' 为图4中 A_4 重构后的数值，相对应 D_1'、D_2'、D_3'、D_4' 分别为图5细节部分各层重构后的数值，则原始值与重构后的误差如图6所示，即用原始值减去重构后的数值。

随后利用得到的重构后的近似部分和细节部分分别建立 $GM(1,1)$ 预测模型，下面将以一组数据进行后验式检验，重构后的近似部分 A_4，即 A_4' 的数据有2000个，由于 $GM(1,1)$ 预测模型在少数据、贫信息的预测中有突出表现，因此首先挑选第1－15个数据，其数据为[5.4562,5.4408,5.2201,5.0545,4.9439,4.8185,4.7481,4.663,4.5631,4.4672,4.4263,4.3706,4.3002,4.2338,4.1526]，根据前10个数据利用 $GM(1,1)$ 模型来预测15个数据，从而得到的预测数据为[5.4562,5.3516,5.2271,5.1054,4.9866,4.8705,4.7572,4.6465,4.5383,4.4327,4.3295,4.2288,4.1304,4.0342,3.9403]，得到的预测的发展系数 a 为0.023549，根据表1可知适合于中短期预测，其灰作用量 u 为5.54342，式(6)中的残差序列 $e^{(0)}(t)$ 为[0,0.00891,－0.00702,－0.05097,－0.04275,－0.05210,－0.00914,0.01648,0.024720,0.03444]，式(7)中的相对误差序列 $\varepsilon(t)$ 为[0%,1.63807%,0.13455%,1.00842%,0.86468%,0.108110%,0.19244%,0.35341%,0.54173%,0.77102%]。其次对重构后的每个细节部分进行累加运算，再利用灰色进行预测，其中重构后的每个细节部分的第1－15个数据如表2所列。

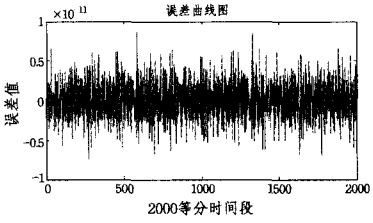

图6　曲线误差图

表2　细节部分各层重构后的数值表

层次	数值
$D4'$	0.16714,0.090868,－0.53756,－1.018,－1.3506,－1.7227,－0.48743,0.31714,0.691,1.1803,0.60705,0.31851,0.31468,0.23456,0.065329
$D3'$	0.16714,0.090868,－0.53756,－1.018,－1.3506,－1.7227,－0.48743,0.31714,0.691,1.1803,0.60705,0.31851,0.31468,0.23456,0.065329
$D2'$	0.62919,0.83814,－1.4447,－3.0599,0.536,2.7356,－0.37674,－2.0657,0.059135,1.1621,0.52443,0.35317,0.28228,0.18449,－0.89666
$D1'$	－0.095994,－0.16627,0.9375,0.19078,－1.9498,1.355,－0.67075,0.32027,－0.0625,－0.54127,1.558,－1.0335,0.05024,－0.77901,2.5245

对表2中的各个重构部分的第1－10个数据进行累加得到的结果为[0.543836,5.59228,－1.22015,－4.05452,－2.94394,2.18161,－1.74809,－1.66295,0.436885,1.53284]，得到的15个数据预测结果为[0.543836,0.54812,－1.28854,－3.94672,－3.15873,1.83756,－1.85421,－1.72561,0.42492,1.60143,2.49518,－0.381257,0.68457,－0.24535,1.86814]，发展系数 a 为－0.0154235，根据表1可知适合于中短期预测，灰作用量 u 为－0.212464，式(6)中残差序列 $e^{(0)}(t)$ 为[0,0.011108,0.06839,－0.1078,0.21479,0.34405,0.10612,0.06266,0.011965,－0.06859,0.07852,0.0106095,0.015186,0.01154,－0.020791]，式(7)中的相对误差序列 $\varepsilon(t)$ 为[0%,1.9863%,5.60504%,2.65876%,7.296%,15.77046%,6.07062%,3.7680%,

(下转第246页)

Learning,1990,5(2):197-227
[14] Robi P. Ensemble learning [EB/OL]. http://www.scholarpedia.org/article/Ensemble_learning,2012-12-11
[15] Buhlmann P. Bagging, Boosting and Ensemble Methods [M]. Berlin:Springer Berlin Heidelberg,2012
[16] Hamid P, Sajad P, Zahra R, et al. CDEBMTE: Creation of Diverse Ensemble Based on Manipulation of Training Example [J]. Pattern Recognition,2012,7329:197-206
[17] 周志华,陈世福. 神经网络集成[J]. 计算机学报,2002,25(1):1-8
[18] Frank A, Asuncion A. UCI Machine Learning Repository [DB/OL]. http://archive.ics.uci.edu/ml. Irvine,CA: University of California,School of Information and Computer Science,2010
[19] Martis R J,Acharya U R,Tan J H,et al. Application of empirical mode decomposition for automated detection of epilepsy using EEG signals[J]. Int J Neural Syst,2012,22(6):1250027
[20] Jayakishan M,Ram B C,Madhab P R,et al. Cascaded Factor Analysis and Wavelet Transform Method for Tumor Classification Using Gene Expression Data[J]. International Journal of Information Technology and Computer Science,2012,4:73-79
[21] Adhvaryu P S,Panchal Mahesh P. A Review on Diverse Ensemble Methods for Classification[J]. IOSR Journal of Computer Engineering,2012,1(4):27-32
[22] Ye Ren,Suganthan P N. Empirical comparison of bagging-based ensemble classifiers [C]// Information Fusion,2012 15th International Conference. 2012:917-924
[23] Tian Jin, Li Ming-qiang, Chen Fu-zan, et al. Coevolutionary learning of neural network ensemble for complex classification tasks[J]. Pattern Recognition,2012,45(4):1373-1385

(上接第 212 页)

2.7387%,4.4747%],最后对两个部分的预测数值进行累加运算,利用第 5 节中第(8)步中的误差进行分析,则得到的结果和误差分析如表 3 和图 7 所示。

表 3 支持度计数向量 SV 实际值与误差值对比表

时间	实际值	预测值	绝对误差	相对误差
1	6	6.00004	0	0%
2	6	5.89972	0.10028	1.67133%
3	4	3.93856	0.061444	1.5361%
4	1	1.15868	−0.15868	15.868%
5	2	1.82787	0.17213	8.6065%
6	7	6.70806	0.29194	4.17057%
7	3	2.90299	0.09701	3.23366%
8	3	2.92089	0.07911	2.637%
9	5	4.96322	0.03678	0.7356%
10	6	6.03413	−0.03413	0.56883%
11	7	6.82468	0.17532	2.50457%
12	4	3.84754	0.152458	3.81145%
13	5	4.81497	0.18503	3.7006%
14	4	3.78885	0.21115	5.27875%
15	6	5.80844	0.19156	3.19266%

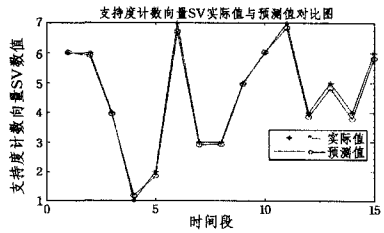

图 7 支持度 SV 实际值与预测值对比图

从表 3 和图 7 可看出预测的精准度比较高。

结束语 利用小波变换多分辨率的特点,把小波变换和灰色模型应用到动态关联规则元规则挖掘中,该方法首先利用小波变换对动态关联规则元规则支持度计数进行处理,这样不仅能保持支持度计数向量在时间上的变化特征,而且能保持在时间上的细节部分,其次利用灰色模型来进行预测,其预测结果从表 3 和图 7 中可以看出预测精准度比较高。然而本文也存在不足之处,如对于时间粒度的划分,现有的动态关联规则的时间段的划分都是等时间段划分,是静态的而动态关联规则的时间粒度的动态划分也是有着变化规律的,同样在动态关联规则元规则挖掘的趋势变化这一块也有待考虑。在文献[15]中论述的趋势度的概念是在支持度和置信度的基础上提出的,因此也可以考虑把趋势度添加进来进行规则的挖掘,这样可以在支持度和置信度的基础上去除无用的关联规则,以对实际应用数据库的关联规则的挖掘做出更好的指导作用。

参 考 文 献

[1] Liu Jin-feng,Rong gang. Mining dynamic association rules in databases[C]// Xi'an Proceedings of International Conference on Computtional Intelligences and Security 2005. Xi'an,2005:688-955
[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):129-133
[3] 沈斌,姚敏. 一种新的动态关联规则及其挖掘算法[J]. 控制与决策,2009,24(9):1310-1315
[4] 刘俊,张忠林,谢彦峰,等. 基于时间序列模型的关联规则元规则挖掘[J]. 计算机工程,2009,15(35):94-96
[5] 胡俊胡,玉清,肖忠卿. 基于小波变换的网络流量预测模型[J]. 计算机工程,2008,34(19):112-114,129
[6] 吴朝阳. 小波变换和 GM-ARMA 组合模型的股指预测[J]. 智能系统学报,2011,6(3):279-282
[7] 白翔宇,叶新铭,蒋海. 基于小波变换与自回归模型的网络流量预测[J]. 计算机科学,2007,34(7):47-54
[8] 佟伟明,李一军,单永正. 基于小波分析的时间序列数据挖掘[J]. 计算机工程,2008,34(1):26-28
[9] Zhang Yi, Wei Yong, Zhou Ping. Improved Approach of Gray Derivative in GM(1,1) Model [J]. The Journal of Grey System, 2006,116(10):160-162
[10] Sun Yan-na. Optimization of Grey Derivative in GM(1,1) Based on the Discrete Exponential Sequence[C]// Proceeding of the 2nd International Symposium on Information Processing(ISTP 2009). Huangshan,P. R. China,2009:313-315
[11] Yang Jiang-tian. Multivariable trend analysis using grey model for machinery condition monitoring[C]// Eleventh World Congress in Mechanism and Machine Science. 2004:2188-2191
[12] 张华,任若恩. 基于小波分解和残差 GM(1,1)—AR 的非平稳时间序列预测[J]. 系统工程理论与实践,2010,30(6):1016-1020
[13] 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,2004:142-146
[14] 刘思峰,邓聚龙. GM(1.1)模型的适用范围[J]. 系统工程理论及实践,2000,20(5):121-124
[15] 张忠林,曾庆飞,许凡. 动态关联规则的趋势度挖掘方法[J]. 计算机应用,2012,32(1):196-198