时间序列模糊关联规则的挖掘

王炳雪
(上海财经大学经济信息管理系,上海 200433)
E-mail:xgdwangbingxue@163.net

摘 要 对于许多复杂系统产生的时间序列,研究序列的局部行为和局部关联特征往往比原来的研究系统全局性模型具有明显的优势。为研究时间序列内部或时间序列间局部形态的关联特征,文章借助模糊集来软化时间序列属性论域的划分边界从而研究时间序列局部形态的模糊关联规则、规则可信度和规则的评价方法。实际算例显示了算法的有效性。

关键词 数据挖掘 时间序列 模糊关联规则

文章编号 1002-8331-(2004)12-0177-03 **文献标识码** A **中图分类号** TP311.13

Fuzzy Assosiation Rules Mining from Time Series

Wang Bingxue
(Dept.of Economics Information Management,Shanghai University of Finance and Economics,Shanghai 200433)

Abstract: On the occasion of dealing with time series hailed from complex system,the investigation of series's local patterns and local relationship has distinct superiority over traditional global models.In order to find rules relating patterns in a time series to other patterns in that series,or patterns in one series to patterns in another series,a fuzzy sub-series discretization method,which soften the effect of sharp boundaries of delegate of each local sub-series,is proposed.Then the local patterns's relationship called fuzzy association rules,rules's confidence and rules's selection measure are studied.The practical calculation shows that the mining of fuzzy assosiation rules is more effective.

Keywords: Data minig,Time series,Fuzzy association rules

1 引言

在经济、技术的很多领域,广泛存在着各种各样的时间序列问题。处理时间序列问题的传统方法是采用以概率论和数理统计为基础的随机过程理论建立线性或非线性模型,进而进行预测或分析[1]。然而,我们能够观察到的往往仅是系统的演化数据,对系统结构和模型参数知之甚少,无法建立合理的数学模型,对系统行为的精确预测效果也难以令人满意[2]。因而近年来人们开始从研究系统的全局行为转为研究系统的局部行为。在解决问题的思路上,由原来纯数学方法,转变为引入模式识别、机器学习、数据挖掘等技术和数学相结合的方法[3,4]。

目前数据挖掘领域对时间序列进行的研究主要限于确定性关联关系的研究。例如,Das[5]等人研究了时间序列局部形态的确定性关联问题,Mark Last[6]等人研究了时间序列形态和后期趋势的关联问题。事实上,时间序列中的形态形式具有不确定性或模糊性,将时间序列形态进行确定性归类和训练是不确切或不合理的,因而有必要引入模糊关联规则的概念。为此,在文献[5][6]研究的基础上,提出能使每一个局部序列软化到代表形态中的模糊离散化处理方法,进而研究反映局部形态关联特征的模糊关联规则、关联规则的可信度和关联规则的评价方法。

2 时间序列模糊离散化处理

效仿文献[5]的方法,首先对时间序列进行离散化处理,但这里并非将时间序列形态进行确定性归类,而是能将每一个局部序列软化到代表形态中。

设 $s=(x_1,x_2,\cdots,x_N)$ 为一时间序列,将一宽度为 w 的时间窗作用于 s 形成一长度为 w 的子序列 $s_i=(x_i,x_{i+1},\cdots,x_{i+w-1})$,将时间窗在时间序列 s 上从始点至终点进行单步滑移,形成一系列宽度为 w 的子序列 $x_1,x_2,\cdots,x_{N-w+1}$,记

$$W(s,w)=\{s_i|i=1,2,\cdots,N-w+1\} \tag{1}$$

为由该时间序列 s 用宽度为 w 的滑窗滑出的子序列集合。

(1)将 $W(s,w)$ 看作为 w 维欧氏空间中的 $(N-w+1)$ 个点,并将它们随机地分到 k 类中,计算每类中心。第 j 类中心第 l 坐标值为:

$$x_{j,l}=\frac{1}{h}\sum_{i=1}^{h}x_{j,l,i},\ l=1,2,\cdots,w;j=1,2,\cdots,k \tag{2}$$

其中,h 表示第 j 类中的子序列数目,$x_{j,l,i}$ 表示第 j 类中第 i 个子序列第 l 坐标值。

(2)以这些中心作为每类的代表点,计算集合 $W(s,w)$ 中每元素 $s_i,i=1,2,\cdots,(N-w+1)$ 属第 j 类代表点的隶属度函数 $\mu_j(s_i)$:

$$\mu_j(s_i)=\frac{\left(\frac{1}{\|s_i-x_j\|^2}\right)^{\frac{1}{b-1}}}{\sum_{c=1}^{k}\left(\frac{1}{\|s_i-x_c\|^2}\right)^{\frac{1}{b-1}}},\ j=1,2,\cdots,k;b>1 \tag{3}$$

其中 $b>1$ 是一个可以控制聚类结果的模糊程度的常数,$|s_i-x_j|^2$ 表示每一点到第 j 类代表点距离的平方。

(3)用当前的隶属度函数更新计算各类中心:

$$x_{j,l}=\frac{\sum_{i=1}^{(N-w+1)}[\mu_j(s_i)]^b x_{j,l,i}}{\sum_{i=1}^{(N-w+1)}[\mu_j(s_i)]^b}\quad j=1,2,\cdots,k;l=1,2,\cdots,w \tag{4}$$

重复以上(2)(3)步的计算,直到各个样本的隶属度稳定。并且将代表点集合记作 $D=\{x_1,x_2,\cdots,x_k\}$,其中 x_j 表示第 j 个代表点。

3 模糊关联规则的挖掘

经过模糊离散化处理后,得到 k 个代表点和各个子序列到每个代表点的隶属度。每个代表点也就是每个代表形态,并且每个子序列到各个代表形态的隶属度之和为 1,$\sum_{i=1}^{k}[\mu_j(s_i)]=1$。定义模糊关联规则的形式为:"如果 A 发生,那么在时间 T 内 B 发生",记作 $A\overset{T}{\Rightarrow}B$,$A,B\in\{x_1,x_2,\cdots,x_k\}$。形态 A 发生的频数 $F(A)$ 定义为:

$$F(A)=\sum_{i=1}^{(N-w+1)}\mu_A(s_i) \tag{5}$$

其中,$\mu_A(s_i)$ 为 s_i 点属第 A 个代表形态的隶属度。

模糊规则 $A\overset{T}{\Rightarrow}B$ 的可信度为:

$$c\{A\overset{T}{\Rightarrow}B\}=\frac{F(A,B,T)}{F(A)} \tag{6}$$

其中 $F(A,B,T)$ 为形态 A 发生后,紧跟着在 T 时间内 B 发生的频数:

$$\begin{aligned}F(A,B,T)&=\left|\{i(s_i\cong A)\wedge(B\widetilde{\in}\{s_{i+w+1},s_{i+w+2},\cdots,s_{i+w+T-1}\})\}\right|\\&=\sum_{i=1}^{(N-w+1)}\mu_A(s_i)\cdot[\mu_B(s_{i+w+1})\vee\mu_B(s_{i+w+2})\vee\cdots\\&\quad\vee\mu_B(s_{i+w+T-1})]\end{aligned} \tag{7}$$

4 有效规则的选择

经过以上的处理后,得到大量具有不同可信度的规则。为了选择最有价值的规则,我们用 Smyth[7]等人提出的 J-measure 方法对所得规则的有效性进行排序。规则 $A\overset{T}{\Rightarrow}B$ 的 J-measure 定义为:

$$\begin{aligned}J(B_T;A)=p(A)\cdot\{&p(B_T|A)\log\left(\frac{p(B_T|A)}{p(B_T)}\right)+\\&[1-p(B_T|A)]\log\left(\frac{1-p(B_T|A)}{1-p(B_T)}\right)\}\end{aligned} \tag{8}$$

其中,$p(A)$ 表示第 A 种形态出现的频率,也就是形态 A 发生的频数 $F(A)$ 和总子序列数之比,$p(A)=\frac{F(A)}{N-w+1}$;$p(B_T)$ 是任意时间窗之后在 T 时间内 B 发生的频率:

$$p(B_T)=\sum_{i=1}^{(N-w+1)}\frac{[\mu_B(s_{i+w+1})\vee\mu_B(s_{i+w+2})\vee\cdots\vee\mu_B(s_{i+w+T-1})]}{N-w+1}$$

即先验概率;$p(B_T|A)$ 是 A 形态出现之后在时间 T 内 B 形态发生的频率,亦即模糊规则 $A\overset{T}{\Rightarrow}B$ 的可信度,即后验概率。直观来看,公式右边的第一部分 $p(A)$ 是希望这种形态出现的次数更多一些,公式右边的第二部分是熵,表示从先验概率 $p(B_T)$ 到后验概率 $p(B_T|A)$ 的信息获得。

5 多维时间序列模糊关联规则的挖掘

在一维时间序列模糊关联规则的基础上,进一步研究多维时间序列模糊关联规则的挖掘。对于 m 维时间序列,经过滑窗处理后得到子序列集合:

$$W(S,w)=\{s_i^h|i=1,2,\cdots,(N-w+1);h=1,2,\cdots,m\}$$

对 $W(S,w)$ 中的 m 个子集:

$$\{s_i^h|i=1,2,\cdots,(N-w+1)\},h=1,2,\cdots,m$$

均用模糊离散化方法处理后,则每个子集均得到 k 个代表点和该集中各个子序列到该集各个代表点的隶属度。每个子集的代表点集合记作 $D^h=\{x_1^h,x_1^h,\cdots,x_k^h\}$,$h=1,2,\cdots,m$,其中 x_j^h 表示第 h 个子集的第 j 个代表点。定义模糊关联规则的形式为:"如果在 V 时间内 A^1 和 A^2 和…和 A^p 和…和 A^h 发生,那么在时间 T 内 B 发生",记作 $A^1\wedge A^2\wedge\cdots\wedge A^p\wedge\cdots\wedge A^h\overset{V,T}{\Rightarrow}B$。其中 $A^p\in D^p,D^p\in\{D^1,D^2\cdots,D^m\},p=1,2,\cdots,h,B\in D^r,D^r\in\{D^1,D^2\cdots,D^m\}$,并且对于 $i,j=1,2,\cdots,h$ 和 $i\neq j$,有 $D^i\cap D^j=\varnothing$。V 时间 A^1 和 A^2 和…和 A^p 和…和 A^h 发生的频数定义为:

$$\begin{aligned}&F(A^1\wedge\cdots\wedge A^h,V)\\&=\left|\{i|A^1\widetilde{\in}\{a_i^1,a_{i+1}^1,\cdots,a_{i+v-1}^1\}\wedge\cdots\wedge A^h\widetilde{\in}\{a_i^h,a_{i+1}^h,\cdots,a_{i+v-1}^h\}\}\right|\\&=\sum_{i=1}^{(N-w+1)}\left\{\left[\mu_{A^1}(s_i^1)\vee\cdots\vee\mu_{A^1}(s_{i+v-1}^1)\right]\cdots\left[\mu_{A^h}(s_i^h)\vee\cdots\vee\mu_{A^h}(s_{i+v-1}^h)\right]\right\}\end{aligned} \tag{9}$$

模糊规则 $A^1\wedge A^2\wedge\cdots\wedge A^p\wedge\cdots\wedge A^h\overset{V,T}{\Rightarrow}B$ 的可信度为:

$$c\{A^1\wedge A^2\wedge\cdots\wedge A^p\wedge\cdots\wedge A^h\overset{V,T}{\Rightarrow}B\}=\frac{F(A^1\wedge\cdots\wedge A^h,V;B,T)}{F(A^1\wedge\cdots\wedge A^h,V)} \tag{10}$$

其中 $F(A^1\wedge\cdots\wedge A^h,V;B,T)$ 为在 V 时间内 A^1 和 A^2 和…和 A^p 和…和 A^h 发生后,紧跟着在 T 时间内 B 发生的频数:

$$\begin{aligned}&F(A^1\wedge\cdots\wedge A^h,V;B,T)=|\{i|A^1\widetilde{\in}\{a_i^1,\cdots,a_{i+v-1}^1\}\wedge\cdots\wedge\\&A^h\widetilde{\in}\{a_i^h,\cdots,a_{i+v-1}^h\}\wedge B\widetilde{\in}\{a_{i+v}^r,\cdots,a_{i+v+T-1}^r\}\}|=\sum_{i=1}^{(N-w+1)}\\&\{\left[\mu_{A^1}(s_i^1)\vee\cdots\vee\mu_{A^1}(s_{i+v-1}^1)\right]\cdots\left[\mu_{A^h}(s_i^h)\vee\cdots\vee\mu_{A^h}(s_{i+v-1}^h)\right]\cdot\\&\left[\mu_B(s_{i+v}^r)\vee\cdots\vee\mu_B(s_{i+v+T-1}^r)\right]\}\end{aligned} \tag{11}$$

多维情况出现的问题是潜在规则成级数增长,为此需采用 Agrawal[8]等人提出的 Apriori 算法首先对非频繁规则进行删剪。

6 参数问题

以上的模糊关联规则挖掘是在滑窗参数 w 和离散类数 k 固定的条件下进行的,w 和 k 的大小直接影响到最终结果,我们的目标是发现有意义的关联规则。一般情况是,研究时间序列的短形态关系时,w 值应取的较短;研究长形态关系时,w 值应取得长一些。分得类数太多,每一类的中的子序列数目太少,不利于计算可信度;分得类数太少,每一类中的子序列的形态相差太远,类中心的代表性太差。一种简单的方法是不考虑参数选取合适与否,而是选取不同 w 和 k 值,一个好的参数应是能够取得具有意义的形态之间的关系。

7 实验

我们选用纽约外汇市场 2003 年 3 月 20 日以前 2000 个交易日日元相对于美元的格林威治时间 22 点中间汇率数据,记价方法以美元标价法。因为对于很小的变化,变量的对数一阶差分近似于该变量的百分比变化,为此我们在研究之前对数据取自然对数后进行一阶差分得到增量序列,然后对增量序列进

行挖掘。

对于以上的数据,首先确定最小频率阈值为2%、最小可信度阈值为50%,针对不同和不同进行挖掘处理,在排除低于阈值的规则后,用J-measure进行排序,排在最前面的四条规则如表1所示,表1中规则的参考图形如图1所示。另外,分别对德国马克、英镑、瑞士法郎、澳大利亚元、法国法郎、加拿大元等币种相对于美元起伏的变化规律也进行了类似研究。

表1 日元对美元时间序列模糊关联规则

w	k	规则	前件支持度(%)	可信度(%)	J-Measure
10	15	$18\overset{10}{\Rightarrow}7$	6.1	63.6	0.0032
15	20	$26\overset{10}{\Rightarrow}12$	5.2	62.8	0.0067
20	25	$15\overset{10}{\Rightarrow}14$	4.7	57.2	0.0073
20	30	$9\overset{10}{\Rightarrow}21$	3.8	59.3	0.0045

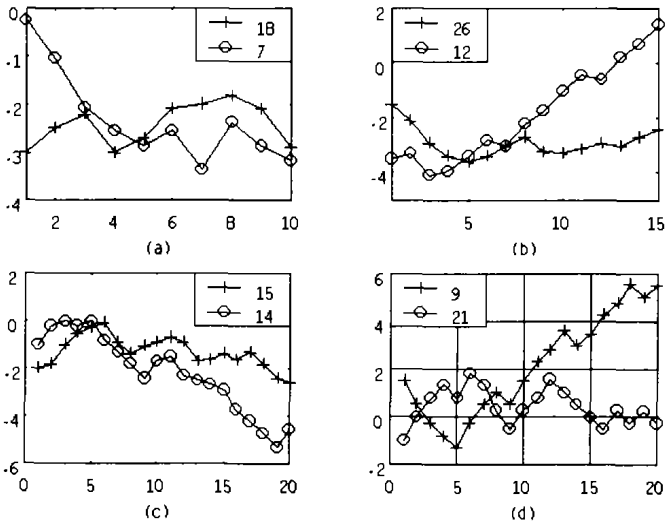

图1 表1中规则的参考图形

为了实验多维模糊关联规则的挖掘,我们选用同样市场、同样时间段、同样记价时间和标价方法的德国马克、日元、英镑、瑞士法郎、澳大利亚元、法国法郎、加拿大元相对于美元的汇率时间序列,试图研究这些序列和日元相对于美元的汇率的关联关系。因为挖掘消耗时间较长,我们仅研究了模糊聚类数目为20,滑窗长度为10的情况,最后得出的模糊规则的可信度最高值为69.1%,比一维情况有所改善。

8 结论

文章借助模糊集来软化时间序列代表形态的划分边界,提出了时间序列数据挖掘的模糊关联规则挖掘方法。实验证明了算法的可行性和优越性。(收稿日期:2003年6月)

参考文献

1.詹姆斯 D 汉密尔顿著.刘明志等译.时间序列分析[M].中国社会科学出版社,1999

2.Andreas S Weigend,Neil A Gershenfeld.Time Series Prediction:Forecasting the future and Understanding the Past[M].Addison Wesley Longman,1994

3.Povinelli R,X Feng.Temporal Patterns Identification of Time Series Data Using Pattern Wavelets and Genetic Algorithms[C].In:Proceedings of Artificial Neural Networks in Engineering,St Louis Missouri,1998:691~696

4.Guralnik V,Srivastava J.Event Detection from Time Series Data[C].In:Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining,San Diego CA USA,1999:33~42

5.Das G,Lin K,Mannila H et al.Rule Discovery from Time Series[C].In:Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining,1998:16~22

6.Last Mark,Yaron Klein,Abraham Kandel.Knowledge Discovery in Time Series Databases[C].In:IEEE Transactions on Systems,Man,and Cybernetics-Part B:Cybernetics,2001

7.Padhraic Smyth,Rodney M Goodman.Rule Induction Using Information Theory[C].In:Gregory P,William J eds.Knowledge Discovery in Databases,Cambridge:the MIT Press,1991:159~176

8.Agrawal R,Mamnila H,Srikant R et al.Fast Discovery of Association Rules[C].In:Fayyad M,Piatetshy-Shapiro G,Smyth P eds.Advances in Knowledge Discovery and Data Mining,Menlo Park,California:AAAI/MIT Press,1996:307~328

(上接95页)

代理的调用句柄实现,来插入到远程方法调用之前

```
public class CommonService implement InvocationHandler
  {
public Object invoke(Object proxy,Method method,Object[],args)
    {
    //implement these common service
    return method.invoke(obj,args);
  }
}
```

通用服务插入到RMI中的调用实现:

```
RemoteObject=(RemoteInterface)Proxy,newProxyInstance
(RemoteObject.getClass.getClassLoader(),new Class[]{RemoteInterface.Class},new CommonService(RemoteObjcet));
```

此时,通过UnicastRemoteObject发布之后,通用服务已经嵌入到动态代理中,当客户请求到来时,先执行嵌入的通用服务,在执行远程方法,可以根据需要,进行通用服务的嵌套,即一种通用服务保持另一种通用服务的Stub,从而可以实现在一个RMI调用中嵌套多种服务。

4 结论

文章利用动态代理实现RMI调用的服务嵌入,保证了RMI实现机制的完整性,同时极大地提高了系统的伸缩性,提高了系统的模块化设计。(收稿日期:2003年6月)

参考文献

1.Eric Gramma,Richard Heml,Ralph Johnson et al.设计模式[M].北京:机械工业出版社,2000

2.Boujarwah A S,Saleh K,AL-Dallal J.Testing syntax and semantic coverage of java language compilers[J].Information and Software Technology,1999;41:15~28

3.Loyd G Williams,Connie U Smith.Performance Evaluation of Software Architecture[C].In:WOSP 98,Santa Fe N M,1998