粗集理论对股票时间序列的知识发现

王晓晔[1,2] 王正欧[1]
[1](天津大学系统工程研究所,天津 300072)
[2](河北工业大学自动化系,天津 300130)

摘 要 提出了将粗集理论应用于时间序列的知识发现。知识发现的过程包括时间序列数据预处理、属性约简和规则抽取三部分。其中数据预处理主要用信号处理技术清洗数据，然后将清洗后的时间序列按照某个变量的变化趋势进行分割,分割后每个时间段内的变化趋势不变,从而将时间序列转换成为一系列静态模式(每种模式代表一种行为趋势),从而去掉其时间依赖性。把决定各种模式的相关属性抽取出来组成一个适用于粗集理论的信息表,然后采用粗集理论对信息表进行属性约简和规则抽取，所得到的规则可以用于预测时间序列在未来的行为。最后将该方法用于股票的趋势预测,取得良好效果。

关键词 知识发现 时间序列 粗集理论 属性约简 规则抽取

文章编号 1002-8331-(2003)29-0099-04 **文献标识码** A **中图分类号** TP18

Knowledge Discovery in Stock Market Time Sieres Based on Rough Set

Wang Xiaoye Wang Zhengou
(Institute of Systems Engineering,Tianjin University,Tianjin 300072)

Abstract: This paper applies Rough set to the knowledge discovery of time series.The process of knowledge discovery in time series includes preprocessing of time series data,attributes reduction and rules extraction.The preprocessing cleans the raw data using the signal processing techniques.Then,the time series is partitioned to a set of pattern(each pattern represents a trend of time series)according to the trend of certain variable.An information table is formed by the most important predicting attributes and target attribute identified from each pattern.This information table is suitable for the Rough set to discover knowledge.Then we use Rough sets to reduce the attributes and extract rules from information table.The extracted rules can predict the time series behavior in the future.We demonstrate our method on time series of stock market data.

Keywords: knowledge discovery,time series,rough set,attributes reduction,rules extraction

1 引言

时间序列数据在科学研究、工程应用和商业交易记录中广泛存在,时间序列数据库与普通数据库相比,有其独特的性质,即时间序列的某些变量是按时间进行排序的,例如股票是一个典型的时间序列,其记录中包括静态变量和动态变量,静态变量是指与时间无关的变量如股票代码、主营项目等,动态变量是指与时间有关的变量如每日的开盘价、收盘价、成交量等。这就使得对时间序列的研究有了特殊性。

近年来人们开始对时间序列进行深入地研究。目前,对时间序列的研究大致集中在以下几方面:

(1)时间序列的相似性研究:一般有两个研究方向,一种是将时间序列从时域映射到频域后再进行相似性匹配,一种是直接在时域内进行研究。主要应用包括:

①从股票数据中识别具有相似变化趋势的模式,以预测新数据的在未来的发展行为。

②超市中具有相似销售模式商品的进货预测等。

(2)时间序列的值预测:包括多步预测和单步预测,将时间序列视为一个动力系统,认为在其过去的波动中蕴涵有可用于预测未来的信息,并以此为基础进行下一步或多步的值预测[1],可用于股票在今后一天或几天的价格预测。

(3)时间序列关联规则的抽取:通过固定长度的窗口将时间序列离散化成一系列子序列，研究子序列之间的相似性,然后将相似的子序列进行聚类形成模式,应用关联规则的研究方法从各种模式中抽取关联规则,可以得到一个时间序列内部不同模式之间的关联规则或不同时间序列之间模式的关联规则[2]。其规则形如“如果第一天 Microsoft 上涨而且 Intel 下降,则 IBM 第二天上涨”。

对时间序列的数据挖掘是从时间序列数据中抽取分类规则。方法是首先通过极值法将时间序列在极值点处分割开,形成一系列模式，则每种模式内部的行为趋势不变（上升过下降),把决定各种模式的条件属性和分类属性组成一个信息表,这个信息表将与时间无关。然后通过粗集理论从信息表中抽取规则形成规则集，用该规则集可以对时间序列进行趋势预测(以股票数据为例，是指用规则集对股票的某种上涨或下跌的行为趋势进行持续时间长或短的预测其规则形式如“如果目前股票形势波动比较厉害,则本次行情持续时间将很短”)。这种预测结果对股票投资者的长期投资行为将更具指导意义。

粗集理论是由波兰科学家 Pawlak 在 1982 年提出的一种

基金项目:国家自然科学基金(编号:60275020)资助;河北省教委基金(编号:401023)资助

处理含糊和不精确问题的新型数学工具[3]。用粗集方法处理不确定问题的最大优点在于它不需要关于数据的预先的或附加的信息，而且容易掌握和使用。其中属性约简和规则抽取是粗集理论在数据挖掘中的一个重要应用。粗集应用于数据挖掘的主要目标是从信息系统表示的数据中抽取规则，而这种数据往往是不依赖于时间的。因此，首先通过上述方法去除时间序列数据的时间依赖性，并形成粗集方法适用的信息表，再用粗集理论从中抽取规则。

该文结构为，第一节介绍时间序列的数据预处理，第二节用粗集理论对处理后的数据进行属性约简和规则抽取，第三节将以上方法应用于股票的行为预测，第四节是总结。

2 时间序列数据库的预处理

时间序列可以看作是通过时间排序的一系列数据，时间序列预处理包括清洗数据、模式分割、属性抽取和离散化四部分。

2.1 数据清洗[4]

在时间序列中可能存在有许多噪声，例如，在股票时间序列中，每天的收盘价包括随机波动和长期趋势数据，因此，在分析这些数据之前必须进行清洗，除去随机波动。

假设原始数据为 $a_{raw}(n)$

$a_{raw}(n)=a(n)+e(n)\quad n=1,2,3,\cdots$

其中 $a(n)$ 为长趋势，$e(n)$ 为噪声，清洗数据应该产生 $\hat{a}(n)$ 来近似描述 $a(n)$。相比较而言，$a(n)$ 是一个稳定的信号，而噪声信号是一个随机的受各种因素影响的信号。如果采用敷立叶变换，可以看到 $a(n)$ 是一个低频信号，$e(n)$ 是一个高频信号。为了滤去噪声，该文采用了信号处理技术中的低通滤波器，其中最简单的一种滤波方法是有限脉冲响应法(FIR)，算式如下

$$\hat{a}(n)=\sum_{i=0}^{N-1}a_{raw}(n-i+[\frac{N}{2}])\cdot c(i)$$

其中

$a_{raw}(n)$ 是原始数据

$\hat{a}(n)$ 清洗后的数据

$c(i)$ 含 N 维系数的向量

N 根据具体数据来定，$c(i)$ 是设计 FIR 的重点，由脉宽和精度来确定，可以用 Matlab 信号处理工具箱中的有关函数来得到。

2.2 模式分割

模式分割的原则是使各个模式具有不同的行为趋势。以股票时间序列为例，收盘价是股票数据中最重要的一个变量，对于广大投资者来说，最关心的是股票行情的持续时间，从而做出买入还是卖出的决定，因此在模式分割时应使每个模式内部收盘价的行为趋势是不变的。模式分割主要是寻找数据中行为趋势改变的转折点[4]。

寻找转折点的最简单的方法是求取曲线的极值点，即 $\left.\frac{d\hat{a}(t)}{dt}\right|_{t=t_x}=0$，由 t_x 组成的一系列时间点 $T_e=\{t_0,\cdots t_{N_e}\}$，将时间序列分割成了 N_e 个模式。

2.3 属性抽取

经过 2.2 节的模式分割，使得每个模式内部数据的行为趋势是固定的，因此可用一个直线方程近似代替原来曲线，则近似直线方程的斜率以及模式区间的长度就是模式的属性。

2.3.1 模式长度

当 $T_e=\{t_0,\cdots t_{N_e}\}$ 中某个区间的宽度 $t_{i+1}-t_i\le d$ 时(d 为设计者设定的阈值)，从 T_e 中除去 t_i 然后插入 $t_i=\frac{t_{i+1}-t_{i-1}}{2}$。

2.3.2 模式斜率

对于上述分割得到的每个区间，求取其近似斜率 α_i

$$\alpha_i=\frac{\hat{a}(t_{i+1})-\hat{a}(t_i)}{t_{i+1}-t_i}$$

则该区间上的近似误差为 β_j $j=1,2,\cdots,N$，N 为 t_i—t_{i+1} 间 $\hat{a}(n)$ 的个数。

$\beta_j=\hat{a}(t_j)-(a(t_i)+\alpha_i\cdot(t_j-t_i))$ 其中 $t_i\le t_j\le t_{i+1}$

2.3.3 信噪比(SNR)

信噪比是时间序列的另一个重要特征，它显示了时间序列的波动情况。信噪比越高证明时间序列越不稳定，受各种因素影响越多。计算信噪比用如下公式

$$SNR_i=\sqrt{\frac{\int_{t_i}^{t_{i+1}}\frac{\varepsilon^2(t)}{\hat{a}^2(t)}}{t_{i+1}-t_i}}$$

式中 $\varepsilon(t)=|a(t)-\hat{a}(t)|$ $a(t)$ 是原始数据

2.4 属性离散化

将上述属性组成信息表用于粗集理论抽取规则时，要求属性值必须用离散(如整型、字符串型或枚举型)数据表达[4]，而显然，上述属性值为连续值，因此必须进行离散化处理，该文采用分箱法离散化，方法如下。

假设属性 $v_1\le a\le v_2$，用二分法将 $[v_1,v_2]$ 分为两个区间 $\left[v_1,\frac{v_1+v_2}{2}\right]$ 和 $\left[\frac{v_1+v_2}{2},v_2\right]$，令 S_1 和 S_2 分别对应于两个区间的样本数。S_{1N} 和 S_{2N} 分别是两个区间的样本在离散化后产生的不一致样本个数，若两个区间都不满足终止条件，则再对上述两个区间进行二等分，如此用递归的方法将各区间进行二等分，直到每个区间都满足终止条件为止，终止条件如下式：

$$\frac{S_{iN}}{S_i}\le\delta \text{ 或者 } S_i\le M$$

下标 i 指第 i 个区间，δ 为最大不一致率，M 为离散区间内样本个数的最小值

3 属性约简和规则抽取

经过第 2 节的数据预处理之后，将原来的时间序列数据转换成一系列与某个时间点不相关的静态模式，从每个模式中抽取出与其相关的属性组成一条记录(或叫做实例)，从而形成了适用于粗集理论的信息表。

粗集理论是由波兰科学家 Pawlak 在 1982 年提出的一种处理含糊和不精确问题的新型数学工具[3]，它为数字逻辑分析、机器学习和模式识别等领域的研究提供了一种新的方法。在粗集理论中，数据约简是一个非常重要的研究课题。在许多系统中，仅有数据库的部分属性是对做出决策是有用的必须保留，消除其他冗余信息是一个必须要做的工作，目前属性约简的方法很多，该文采用的是粗集理论。而对于约简后的数据库进行分析时，其数据量仍然较大，因此必须对信息表进行值约简，即规则抽取，值约简也是基于粗集理论。

下面对上述两项工作分别进行阐述。

3.1 属性约简

基于粗集理论的属性约简是当前的一个研究热点。在已知关于粗集理论的研究成果中，Skowron 提出的可辩识矩阵给属性约简提供了很好的思路。文献[6]在可辩识矩阵的基础上构造了相对差异比较表。

可辩识矩阵的定义如下：

令 $S=(U,A,D)$ 是一个信息系统的信息表，实例集合 $U=\{x_1,x_2,\cdots,x_n\}$ 为论域，$A=\{a_1,a_2,\cdots,a_N\}$ 是条件属性集合，D 是决策属性，$a(x_i)$ 是实例 x_i 在属性 a 上的值，可辨识矩阵的值可表示为下式

$$(c_{ij})=\begin{cases}-1 & D(x_i)=D(x_j)\\ 0 & a(x_i)=a(x_j)\quad D(x_i)\neq D(x_j)\\ 1 & a(x_i)\neq a(x_j)\quad D(x_i)\neq D(x_j)\end{cases}$$

$$i\neq j\quad i,j=1,2,\cdots,n$$

将 S 的任意两个实例进行如上式的比较计算，便可构造一个 $C_n^2=\lfloor(n^2-n)/2\rfloor$ 行 N 列的表 B，N 为条件属性个数，其中行表示原信息表中任意两个实例的对照，列为两个实例的相应属性的按上式的计算结果。将表 B 中所有包含 -1 元素的行删除，剩余的部分为相对差异比较表 $B1$。根据相对差异比较表得出基于粗集理论的属性约简算法。该算法的核心思想是从相对差异比较表 $B1$ 中抽取核属性，从而达到属性约简的目的，具体算法如下：

步骤 1 由信息表 S 构造相对差异比较表 B1、初始化属性约简子集 R=Φ。

步骤 2 分别对 B1 的各列元素纵向相加，结果存入向量 COL。

步骤 3 找出 COL 中最大值所在的列，若只有一列，则该列对应的属性选为 SELECT；如果有 n(n>1)列，则从 n 列中随机选取一列，其对应的属性选为 SELECT。

步骤 4 R←R∪{SELECT}.从 B1 中删除 SELECT 属性列中元素 1 所在的所有行，然后删除 SELECT 列。得到的新表赋给 B1。

步骤 5 如果 B1 非空，转到步骤 2。

步骤 6 R 为信息表 S 的一个属性约简子集。只保留 S 中 R 包含的属性列，删除重复的实例，将新的信息表赋给 S1。

3.2 规则抽取

粗集理论还具有从信息表中抽取规则的能力，实际上，规则抽取的过程正是对信息表进行值约简的过程。该文采用文献[7]的算法，其实质是从信息表的每个实例中寻找对得出决策影响最大的属性。主要步骤如下：

步骤 1 对信息表中的条件属性进行逐列考察.如果删除该属性列后，若产生冲突实例，则保留冲突实例的原该属性值；若为产生冲突但含有重复实例，则将重复实例的该属性值标为“*”；对其它实例，将该属性值标为“？”。

步骤 2 删除可能产生的重复实例，并考察每条标记“？”的实例。若仅由未被标记的属性值即可判断出决策，则将“?”标记为“*”，否则，修改为原属性值；若某个实例的所有条件属性均被标记，则将标有“？”的属性项修改为原属性值。

步骤 3 删除所有条件属性均被标为“*”的实例及可能产生的重复实例。

步骤 4 如果两个实例仅有一个条件属性值不同，且其中一个实例该属性被标为“*”，那么，对该实例如果可由未被标记的属性值判断出决策，则删除另外一个实例；否则，删除该实例。

经过值约简后的信息表形如表 1，c1，c2，c3 为条件属性，class 为决策属性，每个实例代表一条规则，“0”和“1”为属性值，“*”表示在其所在的规则中，其所对应的属性对得出该规则影响很小可以忽略，则每个实例中未被标记为“*”的属性个数为该条规则的条件数，表 1 所代表的规则为：

表 1 值约简后的信息表

U	c1	c2	C3	class
1	1	*	*	0
2	*	0	0	0
3	0	1	*	1

RULE1.If(c1=1)then class=0
RULE2.If(c2=0)^(c3=0)then class=0
RULE3.If(c1=0)^(c2=1)then class=1

4 实验结果(Experiment and Result)

股票数据是一个典型的时间序列数据库，股票数据库的记录包括股票代码，经营项目和每日开盘价和收盘价。其中前两项是静态属性，而每日开盘价和收盘价是动态属性，投资者关注的主要是收盘价的波动，因此本文只对收盘价进行分析。每日收盘价受各种干扰比较多，因此在分析之前必须进行数据清洗。

对于股票投资者来说，最关心的是股票行情的转变，譬如当股票由升变为降时抛出，而由降变为升时买入。因此，如何能够准确预测行情的转折点，是投资者最关心的。这一节，将上面描述的时间序列的知识发现过程应用于上海证券自 1998.1 至 2000.12 三年之间的历史交易数据，分析每个行情的转折点。

4.1 数据预处理

首先采用 2.1 节提到的低通滤波器清洗数据，原始数据是上证的 100 支股票的数据，经过多次实验，脉宽取为 0.1/天，N 取为 30。某支股票经过清洗后的数据见图 1(平滑曲线)。通过 2.2 节的模式分割和 2.3 节的属性抽取，将决定某个模式的属性重组成一系列记录，共得到 864 个记录，每个记录的条件属性包括：前一个模式区间的长度(Length1)，前一个模式区间的斜率(Slope1)，该模式区间的斜率(Slope2)，前一个模式区间的信噪比(Snr1)，本模式区间的信噪比(Snr2)，预测属性为本模式区间的长度(Length2)，即本行情的转折点。因为这些属性都是连续数据，若用于规则抽取，必须进行离散化，采用 2.4 节的分箱法将其离散化，其中 δ 取为 0.2，M 取为 30，离散值的表示方法采用语义表达法，其离散结果见表 2。

表 2 属性离散化结果

属性	离散结果	范围
模式长度(Length)	Short	[0,30]
	Medium	[31,60]
	Long	[61,+∞]
模式斜率(Slope)	Negative- large	$[-\infty,-3*10^{-4}]$
	Negative- little	$[-3*10^{-4},0]$
	Position-little	$[0,3*10^{-4}]$
	Position-large	$[3*10^{-4},+\infty]$
信噪比(Snr)	Low	$[0,4*10^{-5}]$
	Medium	$[4*10^{-5},5*10^{-5}]$
	High	$[5*10^{-5},+\infty]$

4.2 属性约简和规则抽取

将离散化之后的信息表用 3.1 节中提到的方法进行属性约简，发现 Snr2、Slope2 和 Slope1 是核属性，其中 Snr2 贡献最

大,Slope2 次之。由此可见当前区间的信噪比对于转折点的出现起的作用最大。而 Length1 和 Slope1 可以从数据库中删除。最后从约简后的信息表中取出三分之二的数据进行规则抽取,共抽取出 20 条规则,形式如下。

·RULE1:If Snr2 is high then Length2 is short

·RULE2:If Snr2 is medium and Slope2 is position large then Length2 is medium

·RULE3:If Snr2 low and Slope2 is position little then Length is long

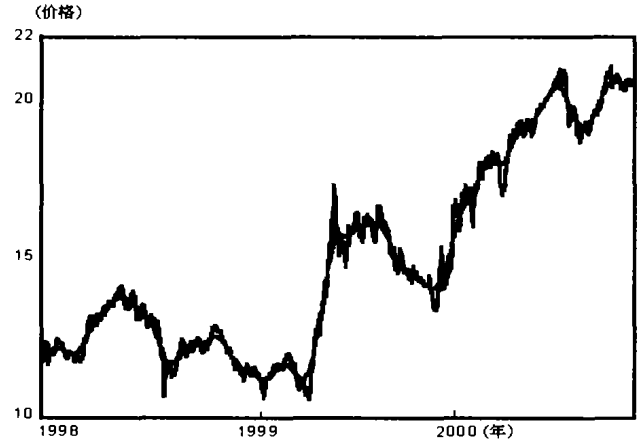

图 1 数据清洗结果

剩余三分之一的数据进行验证,预测精度为 78.5%,而采用 C4.5 对该信息表进行分类,其预测精度仅为 65.5%。分析所得结果,由于粗集理论约简了信息表中的冗余属性和规则中的值约简,因此抽取得到的规则中含有较少的条件属性个数,从而使规则集更加精练且更具概括性,其预测精度相对较高。

需要说明的问题是由于我国股市的发展还处于初级阶段,人为因素和国家政策对股市产生不规范突变,从而大大增加了技术预测的难度,因此技术预测和消息作市相结合将有利于降低股市操作的风险。

5 结论

该文采用粗集理论对时间序列进行知识发现,将时间序列这一与时间密切相关的数据库分割成一系列静态模式,将决定模式的行为趋势的属性组成一个信息表,然后用粗集理论进行模式的行为预测,并把该方法应用于股票的趋势预测,结果表明该方法在处理股票时间序列数据的预测方面具有较好的预测性能,将传统的技术分析方法和现代高科技相结合,将成为股市分析方法发展的必然趋势。(收稿日期:2002 年 10 月)

参考文献

1.杨一文,刘贵忠等.基于小波网络的非线性时间序列预测及其在股市中的应用[J].模式识别与人工智能,2001;14(2)

2.Das G Lin,K Mannila,H Renganathan et al.Rule discovery from time series[C].In:Proc of the 4th International Conference of Knowledge Discovery and Data Mining,New York,16~22

3.Pawlak Z.Rough set[J].International Journal of Computer and Information Sciences,1982;11(5)

4.Last M Klein Y.Knowledge Discovery in Time series Databases[J].IEEE Trans on System,Man,and Cybernetics-part b,2001;31(1)

5.王国胤.Rough 集理论与知识获取[M].西安交通大学出版社,2001:99~104

6.潘丹,郑启伦等.属性约简自寻优算法[J].计算机研究与发展,2001;38(8)

7.常犁云,王国胤等.一种基于 Rough Set 理论的属性约简及规则提取方法[J].软件学报,1999;10(11)

(上接 98 页)

Scenario,基于上述语义标记的信息搜索结果见图 3。图 3 中左边是用户查询上海一历史名园,由查询 agent 提供给用户所搜索到的信息,其中包括了豫园的名字,所在城市,所占面积等;右边是用户查询一个可提供 web 页面搜索服务的名字,查询 agent 返回所搜索到搜索引擎名字"Google"。

4 结论

Q 不是为了描述有关于 agent 内部模型的交互模式,也不是为了描述 agent 间通信和交互的协议,而是为描述用户期望 agent 如何动作的 Scenario,以促进那些非计算机专家,特别是普通的终端用户使用和理解 agent,实现 agent 的社会化。该文介绍了 *Q* 语言的设计目的,设计机理和基本语法构件。通过集成 *Q* 到"语义 Web 上基于 agent 的信息搜索"应用中,笔者利用 *Q* 语言实现了用户对查询 agent 及其所执行 Scenario 的定义,于是从用户的角度控制查询 agent 的动作以实现信息查询任务。应用实例表明 *Q* 语言最重要的特点在于普通用户设计和应用 agent 系统时的简单性,特别突出了非计算机专家的普通用户在利用 *Q* 设计用户和 agent 系统间交互时的优势。
(收稿日期:2003 年 5 月)

参考文献

1.Gerhard Weiss ed.Multiagent Systems-A Modern Approach to Distributed Artificial Intelligence[M].Cambridge,Massachusetts London,England:The MIT Press,1999

2.Ishida T,Fukumoto M.Interaction Design Language Q:The Initial Proposal[J].Transactions of JSAI,2002;17(2):166~169

3.K Kuwabara,T Ishida,N Osato.AgentTalk:Describing Multi-agent Coordination Protocols with Inheritance[C].In:IEEE Conference on Tools with Artificial Intelligence(TAI-95),1995:460~465

4.Nicholas R Jennings,Katia Sycara,Michael Wooldridge.A Roadmap of Agent Research and Development Autonomous Agents and Multi-Agent Systems[J].Journal of Autonomous Agents and Multi-Agent Systems,1998;1(1):7~38

5.T Finin,R Fritzson,D McKay et al.KQML as an Agent Communication Language[C].In:In Proceedings of Third International Conference on Information and Knowledge Management(CIKM'9J)[M].ACM Press,1994

6.R Kent Dybvig.The Scheme Programming Language[M]second edition,Prentice Hall,Inc,1996

7.Service Ontology.http://www.daml.org/daml-s/2001/05/Service.daml

8.DARPA Agent Markup Language+Ontology Inference Layer(DAML+OIL)Specification[S].http://www.daml.org/2001/03/daml+oil-index.html

9.K Sagonas,T Swift,D S Warren et al.The XSB System Version 2.5,Volume 1:Programming's Manual;Volume 2:Libraries and Interfaces[M].2002