基于片段模式的多时间序列关联分析

秦亮曦[1,2,3] 刘新峰[2] 史忠植[1]

(中国科学院计算技术研究所智能信息处理重点实验室 北京 100080)[1]

(中国科学院研究生院 北京 100049)[2] (广西大学计算机与电子信息学院 南宁 530004)[3]

摘 要 本文对基于片断模式的多时间序列关联分析进行了研究,提出了一种分析方法。这一方法是,首先通过聚类找出在时间序列中频繁出现的片断模式,然后将找到的片断模式作为模板,对时间序列进行跨事务关联分析。我们采用中国证券市场1997～2001年的数据为测试数据集,对我们提出的算法进行了测试。测试结果表明,我们的算法是有效的。

关键词 时间序列,关联规则,聚类,动态时间规整

Segment-based Multiple Time Series Association Analysis

QIN Liang-Xi[1,2,3] LIU Xin-Feng[2] SHI Zhong-Zhi[1]

(Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)[1]

(Graduate School of Chinese Academy of Sciences, Beijing 100049)[2]

(College of Computer, Electronics and Information, Guangxi University, Nanning 530004)[3]

Abstract In this paper, it is studied that segment-based multiple time series association rules analysis, and an approach is presented. The approach is that, firstly finding the frequent segment pattern in the time series by means of clustering, then using the found segment patterns as templates to do the inter-transactional association analysis. We use the data of Chinese stock market from 1997 to 2001 as the test data set to test our approach. The experimental results show that the approach is effective.

Keywords Time series, Association rules, Clustering, Dynamic time warping

1 引言

时间序列是由某个物理量在不同时间点的采样值按照时间次序排列而组成的序列。如:股票市场每天的股票收盘价格数据、每天的气温数据、电话公司每小时的话务量等都是时间序列。组成时间序列的数据中,往往隐含着相应系统的一些内在的统计特性和发展规律。如何根据这些时间序列,较准确地找出这些特性和规律,尽可能多地从中提取出我们所需要的信息呢?用以实现上述目的的整个方法称为时间序列分析。这是一种根据动态数据揭示系统动态结构和规律的方法。

对时间序列的分析,已有了许多不同的分析方法,如移动平均法、回归方法、自回归求和移动平均法、神经网络方法等。基于数据挖掘技术的时间序列分析方法,是近些年来发展起来的一种新的研究方法,而基于关联规则的分析方法则是其中的一个研究热点。例如,Lu等人[1]研究了多元跨事务关联规则挖掘问题,他们还提出了基于Apriori思想的跨事务关联规则挖掘算法E-Apriori和EH-Apriori。Das等人[2]研究了将时间序列作为离散的符号序列的关联规则挖掘问题。董泽坤等人[3]也对多元跨事务关联规则挖掘问题做了研究,并提出了分步骤挖掘算法ES-Apriori。秦亮曦等人[4]提出了一种基于压缩FP-树,分而治之挖掘跨事务关联规则的算法IT-ARM。

在时间序列中,往往有许多会不断重复出现的片断模式。如证券市场的股票走势,有许多重要的形态(如头肩型、头肩底、上升三角形、下降三角形、长方形等)。有经验的证券分析人员能够识别出这些形态,并根据历史数据情况判断股票的下一步走势。然而,由于股票数据的动态特征非常强,数据复杂多变,对于大多数投资者而言,仅仅依靠人眼观察,从大量的历史数据中分析和判断股票的走势,是一件非常难办的事情。如果能利用计算机自动地挖掘、分析这些形态以及它们之间的关系,对于投资者预测股票走势将提供非常大的帮助。

本文感兴趣的是发现多时间序列的片断模式之间的先后关联性,即发现形如"在股票A走出形态X、股票B走出形态Y几天之后,股票C一定会走出形态Z。(20%,80%)"之类的规则。

为了研究时间序列中的片断模式之间的关系,首先必须找出这些片断模式。如何在时间序列中发现片断模式,是一项比较重要的研究内容。在模式识别的许多研究中,往往需要借助于领域知识来发现模式。然而在许多情况下,往往无法或无法完整地事先给出要发现的模式的描述,这时依靠算法自动地从时间序列数据中搜索需要的模式是非常需要的。这种算法的优点在于,它既不需要一些有意义的先验条件,也不需要对描述的模式给出详尽的解释。在本文的研究中,我

秦亮曦 副教授,博士研究生,研究方向:数据挖掘、进化计算;**刘新峰** 硕士研究生,研究方向:数据库应用、ERP;**史忠植** 研究员,博士生导师,主要研究方向:知识工程、机器学习、智能主体等。

们不准备借助领域知识定义哪些模式会经常出现，而是采用聚类的方法，从实际的时间序列中找出频繁出现的模式。

本文第2节讨论时间序列的聚类方法以及相似性度量方法；第3节基于片断模式的跨事务关联规则挖掘算法IT-ARM；第4节给出了采用我们的方法对证券时间序列数据进行挖掘的结果；最后是本文的结论。

2 时间序列的聚类

2.1 聚类方法

聚类是根据数据的不同特征，将其分组成分为不同的数据类或簇(Cluster)，使得同一类个体之间的距离尽可能地小，而不同类别个体之间的距离尽可能地大。目前有许多种不同的聚类方法。在本文中，我们使用基于k-中心点的划分方法，对时间序列的频繁片段模式聚类。以下是划分方法的简要描述。

给定一个有 n 个对象或元组的数据集，使用划分方法构建数据的 k 个划分，每个划分表示一个聚簇，并且 $k\leqslant n$。也就是说，它将数据划分为同时满足如下要求的 k 个组：(1)每个组至少包含一个对象，(2)每个对象必须属于且只能属于一个组。划分方法首先创建一个初始划分，然后采用逐步迭代的方法，尝试通过对象在划分间移动来改进划分。聚类时对每个簇需要指定一个簇心。一般有以下两个比较流行的计算簇心的方法：(1)k-均值算法，在该算法中，每个簇用该簇中对象的平均值表示。(2)k-中心点算法，在该算法中，每个簇用接近聚类中心的一个对象来表示。此类方法对中小规模的数据集中发现球状簇很适用。

在对时间序列 $s=(s_1,s_2,\cdots,s_n)$ 进行聚类时，我们定义一个与片断模式长度相等的滑动时间窗口(窗口大小为 w)。将滑动时间窗口应用到序列 s 中，沿着时间的方向，每个窗口内的子序列定义一个标识，然后对于所有的 $n-w+1$ 个子序列聚类。为了降低聚类的难度，我们只需要考虑频繁的片段模式，而非频繁的片断不予考虑。

2.2 相似性度量

在对时间序列的片断模式进行聚类时，必须衡量两个序列的相似程度。相似性度量是时间序列的查询、分类、聚类等许多数据挖掘问题的基础。有许多用于衡量时间序列相似度的方法，其中常用的方法是比较它们之间的“距离”。目前常见的距离计算方法主要有以下三种：

(1) 欧氏(Euclidean)距离：

$$d_{ij}=\sqrt{\sum_{k=1}^{N}(x_{ik}-x_{jk})^2}$$

其中，x_{ik} 和 x_{jk} 分别表示个体 i 和个体 j 的第 k 个属性。

(2) 曼哈顿(Manhattan)距离：

$$d_{ij}=\sum_{k=1}^{N}|x_{ik}-x_{jk}|$$

(3) 明考斯基(Minkowski)距离：

$$d_{ij}=\sqrt[q]{\sum_{k=1}^{N}|x_{ik}-x_{jk}|^q},(q>0)$$

第(3)种方法是第(1)和第(2)两种距离的概化，当 $q=1$ 时，它表示的距离即为曼哈顿距离；而 $q=2$ 时，即为欧氏距离。

除了以上几种距离度量方法之外，还有一些其它的方法。动态时间规整(Dynamic Time Warping；DTW)是一种将时间规整和距离测度结合起来的一种非线性规整技术。这一技术是20世纪60年代由日本学者Itakura提出来的。其主要思想是把未知量均匀地伸长或缩短，直到与参考模式的长度一致。在这一过程中，未知量的时间轴要不均匀地扭曲或弯折，以使其特征与参考模式特征对正。动态时间规整在语音识别领域已经使用了许多年，1994年Berndt等人[5]首先将这一技术应用到数据挖掘领域中。

在时间序列中，由于很多序列形状虽然相似，但往往是不同步的。使用欧氏距离等方法很难识别出这样的相似序列，而使用DTW距离则更适合于识别这样的序列。图1中给出了分别采用欧氏距离和DTW距离计算两个序列的示意图，可见这两个序列有几乎相同的形状，但是它们在表示时间的 X 轴上并没有同步。图1(a)中的欧氏距离计算的是同一时刻两个序列在 Y 轴上的数据的距离；图1(b)中的DTW距离，不再计算相同时刻的 Y 轴距离，取而代之的是那些“非线性‘点-点’对”之间的距离。通过动态时间规整后的距离测度更能反映两个序列之间的相似性。

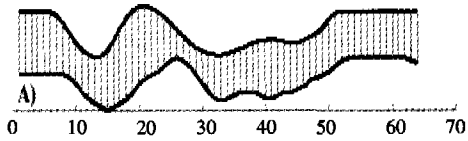

图1 (a) 欧氏距离计算的“点-点”对

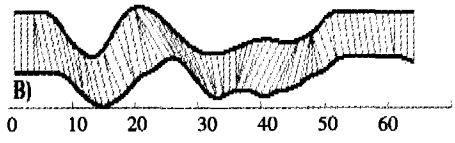

图2 (b) DTW距离计算的“点-点”对

DTW距离是按照如下的过程计算的：

假设我们有两个时间序列 Q 和 C，长度分别是 n 和 m，$Q=q_1,q_2,\cdots,q_i,\cdots,q_n$；$C=c_1,c_2,\cdots,c_j,\cdots,c_m$。我们构造一个 (n,m) 的矩阵，第 (i,j) 单元记录两个点 q_i 和 c_j 之间的欧氏距离 $d(q_i,c_j)=|q_i-c_j|$。一条弯折的路径 W，由若干个彼此相连的矩阵单元构成，这条路径描述了 Q 和 C 之间的一种映射。设第 k 个单元定义为 $w_k=(i,j)_k$，则

$$w=w_1,w_2,\cdots,w_k\cdots,w_K\quad \max(m,n)\leqslant K\leqslant m+n-1$$

这条弯折的路径满足如下的条件：

(1)边界条件：$w_1=(1,1)$ 且 $w_k=(n,m)$。也就是说，这条路径必须由第一个单元经过矩阵到达最后一个单元。

(2)连续性：设 $w_k=(a,b)$，$w_{k-1}=(a',b')$，那么，$a-a'\leqslant 1$，$b-b'\leqslant 1$。这个条件限制了允许的每一个弯折步必须彼此相邻(包括对角线单元)。

(3)单调性：设 $w_k=(a,b)$，$w_{k-1}=(a',b')$，那么，$a-a'\geqslant 0$，$b-b'\geqslant 0$。这个条件限制了 W 中的点在时间上的不能回退。

在满足上述条件的许多路径中，最短的、花费最少的一条路经是：

$$DWT(Q,C)=\min\{\sqrt{\textstyle\sum_{k=1}^{K}w_k}/K\}$$

路径上的点 w_k 对应的 q_i 和 c_j 即为DTW距离的“点-点”对。

DTW 距离的计算过程是一个逐步迭代的过程。首先初始化矩阵的每一个单元的局部距离 $d(i,j)=|q_i-c_j|$，然后通过如下公式：

$$\gamma(i,j)=d(i,j)+\min\{\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\}$$

迭代地计算累计距离 $\gamma(i,j)$。当所有的点都计算完毕时，就得到了两个序列的 DTW 距离。

3 频繁时间序列片段的关联分析

在证券市场中，不同的股票之间往往存在着关联联动性。我们在通过采用聚类方法从数据中找出频繁片断模式之后，就可以采用关联规则的分析方法来分析不同序列的片断模式之间的关联性。我们采用多时间序列的跨事务关联规则挖掘算法 ITARM[4] 对片段模式进行跨事务关联规则分析。

ITARM 算法主要是基于 CFPmine 算法[6] 的思想，并对挖掘任务采用了进一步分解的方法。在产生了频繁 1-项集之后，分别利用 1-项集中的项作为约束条件，得到在某一项约束条件下的频繁 1-项集。然后利用这些约束下的频繁 1-项集建立压缩 FP-树，采用 FPmine 算法挖掘压缩 FP-树，得到约束下的所有频繁集或生成关联规则。将所有这些约束下的频繁集或关联规则合并就可得到完全集。具体实现思想如下：

1) ITARM 算法将多时间序列中的所有 1-项集的数据库信息读入内存，保存于一个字节数组中。这样，后面所有频繁项集的支持度计数都不必搜索数据库，从而减少了系统的 I/O 代价。如果数据量非常大，也可以考虑将概化后的数据保存在文件中，不过相对要增加一些读数据的时间。

2) 将挖掘分为两个步骤：第一步找出所有满足最小支持度阈值的事务内频繁 1-项集；第二步，在此基础上挖掘跨事务序列的频繁项集和关联规则。

3) 第二步的挖掘采用分而治之的方法。对于每个满足条件的参考时间基准点项 $e_i(0)$，执行如下的操作：

(1)首先找出在 $e_i(0)$ 出现的情况下，滑动窗口内的频繁 1-项集 $F1_i$（若加上 $e_i(0)$，则相当于频繁 2-项集）。

(2)将 $F1_i$ 按照 $(e_{i+1}(0),\cdots,e_u(0),e_1(1),\cdots,e_u(1),\cdots,e_1(w-1),\cdots,e_u(w-1))$ 的次序排列，设排序后的 $F1_i$ 为 $OF1_i$。

(3)扫描已读入内存的数据集 D，将每个滑动窗口作为一个事务，找出 $e_i(0)$ 出现时该事务内的频繁项，构造一棵压缩 FP-树(CFP-树)[6]。CFP-树与 FP-树的区别主要有：①FP-树的每个结点都包含 6 个域，而 CFP-树每个结点只包含 4 个域，因此 CFP-树所占用的内存空间仅为 FP-树的 2/3。②FP-树中的结点是无序的，而 CFP-树中的结点是按照 item-no 从小到大的次序排列的。③FP-树是双向的，而 CFP-树是单向的。树建立时只有从树根到叶结点的路径，树建立后只存在从叶结点到树根的路径。

(4)对 CFP-树的挖掘是调用 CFPmine 过程完成的。当某一项挖掘完成后，立即输出以 $e_i(0)$ 为起始的频繁项或关联规则，然后删除该树，再进入下一项的挖掘。

算法 1 ITARM 算法

输入：时间序列合并集 D，最小支持度阈值 min_sup，最小可信度 min_conf，滑动时间窗口 w
输出：D 中的跨事务关联规则集
方法：
Phase 1
(1) $C_1=\{\{e_i(x)\}\mid (e_i(x)\in\Sigma)\wedge(0\leqslant x\leqslant w-1)\}$
(2) for each inter-time series transaction T_x in D
(3) for each candidate c: $e_i(x)\in C_1(e_i(x)\in T_{i+x})$
(4) c. count ++;
(5) $L_1=\{c\mid\{e_i(x)\}\mid(c\in C_1)\wedge(c.\,count\geqslant support)\}$
Phase 2
(6) for each item: $e_i(0)\in L_1$ {
(7) $C'_2=\{\{e_i(0),e_k(x)\}\mid e_k(x)\in L_1((x\neq 0)\vee(x=0\wedge i<k))\}$
(8) for each candidate $c\in C'_2$: $\{e_i(0),e_k(x)\}$ {
(9) c. count ++; }
(10) $L'_2=\{c\mid\{e_i(0),e_k(x)\}\mid(c\in C'_2)\wedge(c.\,count\geqslant min\_sup)\}$ // L'_2 即为 $F1_i$
(11) 将 L'_2 排序
(12) 扫描数据集 D，建立 CFP-tree
(13) 调用 CFP-树挖掘算法 CFPmine，得到以为 $e_i(0)$ 起始的所有频繁项集
(14) 利用频繁项集生成并输出关联规则
(15) 删除 CFP-tree
(16) }

4 实验及结果

为了验证算法的有效性，我们使用中国证券市场近 500 只股票 1997～2001 年的收盘价数据为测试集，对股票数据时间序列进行片断模式聚类，并在此基础上进行基于片断模式的跨事务关联分析。经过数据预处理、序列片段模式聚类、关联规则挖掘等处理过程，我们发现了一些有趣的“知识”。以下是实验的步骤及部分结果。

4.1 平滑

实际的股票收盘价格构成的序列是非平稳的，序列与序列之间差异很大。为了能够容易地对它们进行分析，需要将每个序列差分后再处理，这样序列就转换为平稳序列。

然而，转换后的平稳序列扰动特征明显，必须再经过平滑处理。否则，序列中的突发事件会在 DTW 距离计算中跳过而丢失它携带的信息。一般较多用到的平滑处理技术是简单移动平均线：

设观测序列为 $y_1,y_2,\cdots,y_t$，正整数 $N<t$。一次移动平均值 M_t 计算公式为：

$$M_t=\frac{1}{N}(y_t+y_{t-1}+\cdots+y_{t-N+1})$$

公式中的 N 值需要在实际使用中调整，太小的值会失去使用意义，太大的值会损失大量细节。在试验中，我们设定 $N=20$。

4.2 正则化

时间序列中采样值的统计分步不可能都在同一水平附近波动，必须经过正则化处理，把它们放在同样的幅值内比较。

设输入序列的最大值为 I_{max}，最小值为 I_{min}，缩放后的序列最大值为 V_{max}，最小值为 V_{min}，输入序列的每一个值 V 经过处理后输出的值 I 为：

$$I=I_{min}+(I_{max}-I_{min})*(V-V_{min})/(V_{max}-V_{min})$$

4.3 片段模式聚类结果

经过平滑处理、正则化处理后，就可以对时间序列片段模式聚类，进而挖掘频繁片段模式之间的关联规则。在试验中，我们使用多个时间序列为样本集，使用 DTW 距离计算子序列之间的相似度，从中找出一些频繁的片段模式。

这些片段是经过一阶差分处理的 20 日移动平均线中 5 个连续观察值的子序列。这些序列经过正则化处理后，观察值最大为 5，最小为 0。在聚类过程中，最小 DTW 距离为 1.5。如果每个片断的相似片段数超过总片段数的 2%，则认为它是频繁的。图 2 列出了从股票时间序列中发现的部分频繁片段模式。

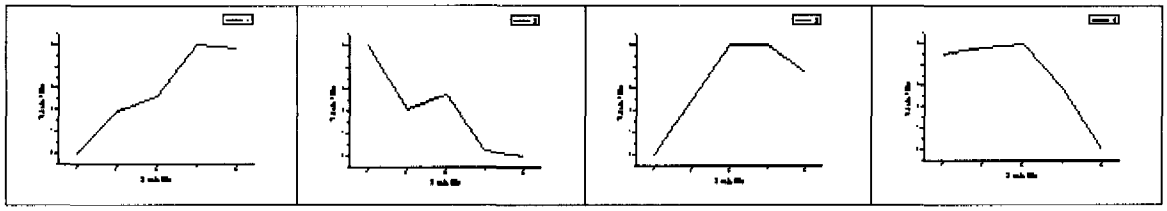

图 2　股票时间序列的部分频繁片段模式

4.4　获得的规则

我们以聚类得到的频繁片段模式作为模板，设定 IT-ARM 算法的参数：最小支持度 0.5%、最小信任度 60%、滑动时间窗口 3、规则长度 3，发现了许多关联规则。以下是部分规则。

R1：宏盛科技第 0 天是模式 11⇒青海三普第 13 天是模式 2(0.5%，83.3%)

R2：云维科技和友好集团第 0 天是模式 0⇒ST 轻骑第 1 天是模式 0(1%，90%)

R3：青岛啤酒第 0 天是模式 0，上菱电器第 0 天是模式 7⇒交运股份第 12 天是模式 0(0.5%，83.3)

结论　我们对基于片断模式的多时间序列关联分析进行了研究，提出了一种分析方法。这一方法是，首先通过聚类找出在时间序列中频繁出现的片断模式，然后将找到的片断模式作为模板，对时间序列进行跨事务关联分析。我们采用中国证券市场 1997～2001 年的数据为测试数据集，对我们提出的算法进行了测试。测试结果表明，我们的算法是有效的，它能发现多个股票之间的频繁片断模式的相互关联性，并以此预测一些股票的未来走势。

参 考 文 献

1　Lu H, Feng L, Han J. Beyond Intra-Transaction Association Analysis: Mining Multi-Dimensional Inter-Transaction association rules. ACM Transactions on Information Systems, 2000, 18(4): 423～454

2　Das G, Lin K, Mannila H, et al. Rule discovery from time series. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). AAAI Press, 1998. http://citeseer.ist.psu.edu/das98rule.html

3　董泽坤，李辉，史忠植．多元时间序列中跨事务关联规则分析的高效处理算法．计算机科学，2004，31(3)：108～111

4　秦亮曦，史忠植．多时间序列跨事务关联分析研究(已投《软件学报》)

5　Berndt D, Clifford J. Using dynamic time warping to find patterns in time series. In: Working Notes of the Knowledge Discovery in Databases Workshop, 1994. 359～370

6　Qin L, Luo P, Shi Z. Efficiently mining frequent itemsets with compact FP-tree. In: Shi Z, He Q, eds. Proc. of Int'l Conf. on Intelligent Information Processing 2004(IIP2004), Beijing, China Springer Press, 2004. 397～406

(上接第 228 页)

想。如实验部分所述，本文方法能够建立更优的语言模型。

在研究对象和实验方面，文[1]和文[2]都只在单字节的一些西方语言字符编码方案上进行了实验。实际上，根据我们的调研，尚没有任何文献发表在中文、日文等双字节编码的语言上的语种识别实验结果。本文重点实验了针对 4 种双字节编码方案 GB2312、BIG5、Shift-JIS 和 eucKR 的语种识别效果；同时，在实验中对比研究了英、俄、法、德等 4 种单字节编码语种。

还有一些面向其他应用(如语音识别)的研究工作讨论如何进一步改善语言模型，如文[6,7]等。从实验结果看，我们认为更复杂的语言模型对于语种识别任务并不是必要的，孤立地追求训练过程的复杂化甚至可能对时间性能带来负面影响。

结论　本文研究了基于字符层马尔科夫模型的语种识别方法。有针对性的考虑了前人工作未曾很好解决的汉、日等双字节编码、词间无空格分隔的语种的识别问题，采用了一种基于 EM 算法的统计语言模型训练算法，同时也对解码算法进行了改进。本文工作的特点不仅体现在引入了基于 EM 算法的模型参数估计能够得到更为优化的语言模型，还体现在对于前人未讨论过的中、日等双字节编码、无空格分隔的语种的识别取得了良好实验结果。

基于字符层马尔科夫模型的语种识别方法具有较好的实用性，为我们的多语机器翻译系统[5]和多语言信息抽取研究提供了有益支持。但当前工作也存在些需要进一步改进的地方。后继工作将更多地面向算法的实用化问题，包括阈值参数 C_0 的自动选择、识别算法时间性能的改进等。

参 考 文 献

1　Cavnar W B, Trenkle J M. N-gram based text categorization. In 1994 Symposium on Document Analysis and Information Retrieval in Las Vegas, 1994

2　Ted D. Statistical Identification of Language: [Technical report CRL MCCS-94-273]. Computing Research Lab, New Mexico State University, 1994

3　Jelinek F, Mercer R L. Interpolated estimation of Markov source parameters from sparse data. In: Proc. of the Workshop on Pattern Recognition in Practice, Amsterdam, The Netherlands: North-Holland, 1980

4　Dempster A, Laird N, Rubin D. Maximum-likelihood from Incomplete Data via the EM algorithm. J. Royal Statist. Soc. Ser. B., 1977(39): 278～286

5　黄河燕，陈肇雄．基于多策略的交互式智能辅助翻译平台总体设计．计算机研究与发展，2004，41(7)：1266～1272

6　Goodman J T. A Bit of Progress in Language Modeling Extended Version: [Technical Report MSR-TR-2001-72]. Microsoft Research, Redmond, July 2004

7　Chelba C. Exploiting Syntactic Structure for Natural Language Modeling: [Phd Thesis]. Johns Hopkins University, 2004