基于小波分析的时间序列数据挖掘模型

郑 诚[1,2] 蔡庆生[3]
[1](安徽大学计算机科学与技术学院,合肥 230039)
[2](安徽大学计算智能与信号处理教育部重点实验室,合肥 230039)
[3](中国科学技术大学计算机科学与技术系,合肥 230027)
E-mail:zhench@ahu.edu.cn

摘 要 论文提出一个基于小波分析的时间序列挖掘模型TSMiner,它支持时间序列数据挖掘的整个过程。该模型由5部分组成:原始数据的可视化、数据预处理、数据约简,模式发现和结果模式可视化。该模型应用小波实现数据的多层次可视化表示、数据约简和多尺度模式发现。它可以帮助用户观察高维数据,理解中间结果和解释发现的模式。

关键词 时间序列 数据挖掘 可视化 小波分析

文章编号 1002-8331-(2004)30-0026-03 **文献标识码** A **中图分类号** TP311

A Model Based on Wavelet Analysis for Data Mining on Time Series

Zheng Cheng[1,2] Cai Qingsheng[3]
[1](School of Computer Science and Technology, Anhui University, Hefei 230039)
[2](Educational Department Key Laboratory of Intelligent Computing & Signal Processing, Anhui University, Hefei 230039)
[3](Department of Computer Science and Technology, University of Science and Technology of China, Hefei 230022)

Abstract: TSMiner, a model for time series visual data mining based on wavelet has been prosposed. The model consists of five components: original data visualization, data preprocess, data reduction, pattern discovery and pattern visualization. By wavelets the model performs hierarchical representation of time series dataset for visualization, data reduction and multi-scale pattern discovery. This model can help users view the high dimensional data, understand the intermediate results, and interpret the discovered patterns.

Keywords: time series, data mining, visualization, wavelet analysis

1 引言

近几年来可视化数据挖掘也已成为数据挖掘研究方向新的研究热点[1-3]。可视化数据挖掘的目标是提供可视化和数据挖掘的结合,以增强整个数据挖掘循环过程的有效性。研究成果表明可视化数据挖掘技术在知识发现中具有很高的价值和潜力。

论文工作在于利用小波分析的多分辨能力,结合数据挖掘技术和可视化技术,构造一个基于小波分析的时间序列数据挖掘模型TSMiner(Time series Miner)。

2 相关工作

2.1 时间序列数据挖掘

数据挖掘领域研究者已提出许多从时间序列中提取有趣模式的新技术。如:相似模式发现技术有:动态时间弯曲[4]、离散付立叶变换(DFT)[5]和小波变换(DWT)[6],规则发现技术[7],分类和聚类技术[8]。

2.2 信息可视化

传统的可视化技术有散点图、直方图、圆饼图、折线图等。近几年来,已提出多种适用于多维数据挖掘的可视化方法[9],如基于像素的方法,几何投影方法、基于图标的方法、层次和基于图形的方法等。

2.3 小波变换

设$\{V_j\}$是一给定的多分辨分析[10],φ和ψ分别是相应的尺度函数和小波函数。

可以用V_j上的基函数来表示尺度函数$\varphi(x)$和小波函数$\psi(x)$:

$$\varphi(x)=\sum_k h_k\varphi(x-k)$$

$$\psi(x)=\sum_k g_k\varphi(x-k)$$

依据多分辨分析,对于任意的j,这些关系在V_{j-1},V_j和W_j之间也是有效的。称h_k和g_k为滤波器系数,它们唯一地定义了尺度函数$\varphi(x)$和小波函数$\psi(x)$。

而$V_{j-1}=V_j\oplus W_j$,可以将一个用V_{j-1}的函数基表示的函数

基金项目:国家自然科学基金项目(编号:60273043);安徽省教育厅自然科学研究项目(编号:2002kj009)

作者简介:郑诚,男,博士,副教授,主要研究领域为数据挖掘与知识发现,人工智能及其应用,数据库及其应用。蔡庆生(1938-),男,教授,博士生导师,主要研究领域为数据挖掘与知识发现,人工智能及其应用。

$f(x)$用 V_j 和 W_j 的函数基来表示：

$$f(x)=\sum_n C_n^{j-1}\varphi_{j-1,n}(x)=\sum_k C_k^j\varphi_{j,k}(x)+\sum_k D_k^j\psi_{j,k}(x)$$

变换系数 C_k^j 和 D_k^j 定义如下：

$$C_k^j=\sum_n C_n^{j-1}h_{n-2k} \quad (1)$$

$$D_k^j=\sum_n C_n^{j-1}g_{n-2k} \quad (2)$$

称 C_k^j 和 D_k^j 分别为分辨率 2^j 下的离散逼近和细节。

用小波分析提取低频信号，低频信号是原始数据的逼近，变换一次后，数据的个数缩小一半，同时新的序列又能很好地保持原序列的趋势特征。且这种变换可以连续进行多次。

3 TSMiner 模型

3.1 功能

(1)原始数据的可视化

通过原始数据的可视化这个过程可以看出数据是如何分布的。以帮助确定序列的分布，走势，找出缺省值。可以采用小波进行多层次逼近表示。

(2)数据预处理

数据预处理是一个操作序列，它将原始数据转换成数据挖掘算法的目标格式。它包括几个方面，如，数据标准化，处理缺省数据值和噪声数据，数值的离散化等。

(3)数据约简

去除一些冗余数据和无意义数据，或由于时间复杂性或内存的能力，选择一小部分数据表示整个集合，对原始数据变换和特征提取。TSMiner 模型采用小波分析进行数据约简。

(4)模式发现

用挖掘算法从大的数据集中找出模式，可借助可视化技术，调整挖掘过程的阈值和参数。可视化约简后的曲线的形状、颜色、亮度等，也可以提供用于发现模式的信息。

(5)模式可视化

将发现的模式在屏幕上可视化输出。用户通过观察结果，可以分析模式的意义。

3.2 结构

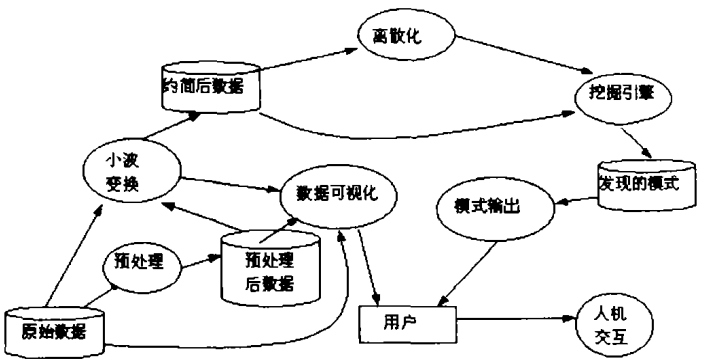

图 1 TSMiner 模型的数据流图

图 1 中表示模型的数据流图。整个挖掘流程中，用户可以通过交互接口控制所进行的工作。

3.3 挖掘引擎

TSMiner 中挖掘引擎包括：(1)相似模式发现：基于编辑距离的相似模式发现；(2)形状相似发现；(3)聚类分析；(4)分类分析；(5)关联分析；(6)趋势分析；(7)特定模式发现。挖掘引擎组成如图 2 所示。

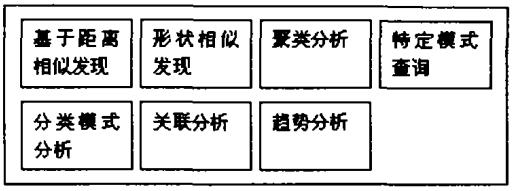

图 2 TSMiner 挖掘引擎组成

4 可视化技术

用小波分析对数据进行多尺度逼近表示，参见图 3 所示。这种方式在时间序列数据挖掘中更为直观。

图 3 基于小波变换的序列多分辨逼近可视化表示

小波变换是可逆的，且适合数据的分层表示。低分辨率概览数据，高分辨率查看包含细节的数据。

X 是一个时间序列数据，可以将它分解为两部分，一部分为逼近部分 AX，另一部分为细节信号 DX，Ha 、Ga 分别为低通和高通滤波器系数，Conv 为卷积操作。将公式(1)(2)转换为下列算法。

```
Begin
    Temp-AX=Conv(X,Ha);
    Temp-DX=Conv(X,Ga);
    AX=Down-sample(Temp,AX)
    DX=Down-sample(Temp,DX)
End;
    (AX,DX)=Descomp(X)
```

需要访问高分辨数据时，能重建序列。小波提供独特的能力建立多分辨层次表示，中间层次可以从低层重构。如果任何一个阶段的逼近和细节信号可以得到，就可以得到高一分辨层的逼近部分。例如，第 I+1 步信号逼近和细节可以得到，第 I 步的逼近可以通过它们重构得到。

5 原型构成

5.1 基本功能

TSMiner 分成三个区域，顶部的主菜单和工具栏，中间是数据序列和模式的可视化输出区，底部是小波选择、小波图形和滤波器系数显示区。

文件菜单具有读取数据、存盘、打印等功能。预处理菜单实现对数据预处理。挖掘引擎目前只集成了一个多尺度相似模式匹配算法和特定模式查询算法。控制菜单设置挖掘的参数，阈值的输入等。选项可以设置数据的显示格式、颜色设置等。Help 提供帮助功能。主界面参见图 4 所示。

5.2 交互接口

数据源的连接接口：打开数据文件并显示。这里可以选择数据格式和数据显示的模式。在不知数据格式的情况下，尝试以多种方式打开，在格式已知的情形下，可以选择数据格式。数

据显示的模式可以是表格、直方图、折线图。

数据预处理接口:通过接口,用户可以根据情况,处理缺省值或某些特定的数值。如股票交易中,经常出现股票停牌情况,停牌期间,行情数据库中价格数据会缺省或值为0,预处理时,停牌期间价格,取前一天的收盘价、或取前一天的收盘价和恢复交易的收盘价的均值。

数据的选取:用户从数据中选择一个感兴趣的序列或子序列,以它作为模板或例子,从数据集中找出与它相似的数据序列。系统中提供"基于例子的查询"的功能,用户可以用鼠标点击数据显示窗口,把它拖入到查询窗口,释放鼠标,这样形成一个查询。查询结果窗口中,找到查询结果的曲线可以用颜色和粗细帮助增强表示。

小波选择接口:选择小波,选择尺度(单个尺度,几个尺度),为用户比较在不同的小波下,数据挖掘的结果,可以从中找出满意的结果。系统中集成了Daubechies小波系的db2、db3、db4等。

参数和阈值的输入:当用户不满意结果时,他可以重新设置参数和阈值,继续搜索,重复这一过程直到用户满意为止。

选项中可以设置数据的显示格式和颜色,对于匹配或未匹配的曲线用不同颜色显示。

5.3 举例

实验数据选用上海和深圳证券交易所的股票日收盘价格数据集。

图4给出了实验中,从股票行情数据库中搜索出的具有相似股价运动趋势的两支股票的情况。最上面两条曲线分别是这两股票价格的原始数据。中间部分别为原始数据经小波在5个尺度变换下的逼近曲线。中间左边为第一支股票变换情况,右边为第二支股票变换情况。从图中可以清楚看到,随着观察尺度的变大,两支股票呈现出越来越相似的特征。

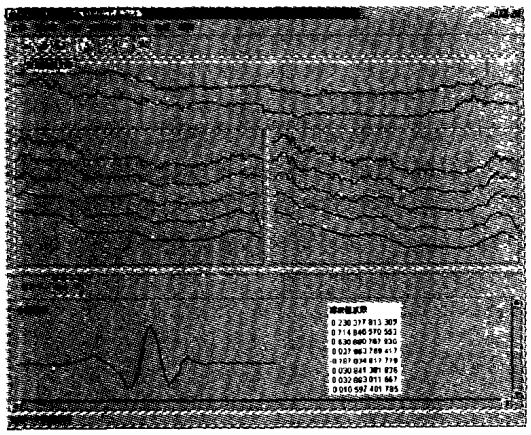

图4 两支股票收盘价序列及它们在5个尺度上相似比较

实验中采用了小波db4,图4左下边给出它的图形,右下边给出它的低通滤波器系数。

6 结论及进一步工作

针对时间序列数据具有高维性的特点,TSMiner模型以小波分析作为基础,集成了可视化和交互式技术与数据挖掘技术,它支持数据挖掘的整个过程。这个过程中,小波分析既解决了高维数据的可视化问题,又可以为数据挖掘解决数据约简问题。人能直观观察数据和结果模式,并通过接口与系统进行必要的交互,能获得更易理解的结果,并能加速数据挖掘的过程。

进一步工作包括,集成其他的时间序列数据挖掘技术,如规则发现、分类和聚类等,完善整个系统。

(收稿日期:2004年7月)

参考文献

1.Ankerst M,Elsen C,Ester M et al.Visual Classification:An Interactive Approach to Decision Tree Construction[C].In:Proc 5th Int Conf on Knowledge Discovery and Data Mining,San Diego,CA,1999:392~396

2.Mihael Ankerst.Visual Data Mining with Pixel-oriented Visualization Techniques[C].In:ACM SIGKDD Workshop on Visual Data Mining,San Francisco,CA,2001

3.Heike Hofmann,Arno P J,Siebes M et al.Visualizing association rules with interactive mosaic plots[C].In:KDD 2000,2000:227~235

4.Berndt DJ,Clifford J.Using dynamic time warping to find pattern in time series[C].In:AAAI Workshop on Knowledge Discovery in Database,KDD-94,Seattle,Washington,1994

5.Agrawal Rakesh,Faloutsos Christos,Swami Arun.Efficient Similarity Search In Sequence Databases[C].In:Proc of the 4th Conference on Foundations of Data Organization and Algorithms,Chicago,1993-08:69~84

6.Chan Franky,Fu Wai-chee.Efficient Time Series Matching by Wavelets[C].In:15th IEEE International Conference on Data Engineering,Sydney,Australia,1999:126~133

7.Gautam Das,King-Ip Lin,Heikki Manilla et al.Rule Discovery in Time Series Databases[C].In:Proc of the Fourth International Conference on Knowledge Discovery & Data Mining,1998:16~22

8.Keogh E,Pazzani M.An enhanced representation of time series which allows fast and accurate classification,clustering and relevance feedback[C].In:Proceedings of the Fourth International Conference of Knowledge Discovery and Data Mining,AAAI Press,1998:239~241

9.Keim DA,Kriegel H-P.Visualization Techniques for Mining Large Databases:A comparison[J].IEEE Transactions on Knowledge and Data Engineering,1996;8(6):923~938

10.Mallat Stephane.A Theory for Multi-resolution Signal Decomposition:The Wavelet Representation[J].IEEE Transactions on Pattern analysis and Machine Intelligence,1989;11(7):674~693

(上接7页)

2.王国俊.非经典逻辑与近似推理[M].北京:科学出版社,2000

3.E P Klement,R Mesiar,E Pap.Triangular Norms[M].Kluwer Academic Publishers,2000

4.E P Klement,R Mesiar,E Pap.Triangular Norms.Position Paper I:basic analytical and algebraic properties[J].Fuzzy Set and Systems,2004;(143):5~26

5.Hajek P.Matamathematics of Fuzzy Logic[M].Kluwer Academic Publishers,1998

6.Stanley Burris,H P Sankappanavar.A Course in Universal Algebra[M].Springer-verlag New York Heidelberg Berlin,1981