

基于有向图构造的数据流预测算法

由 欣, 印桂生

(哈尔滨工程大学 计算机与科学技术学院, 黑龙江 哈尔滨 150001)

摘 要: 数据流的预测技术在金融证券管理、网络监控等应用中具有重要的作用. 分析近年来有关数据流预测的技术, 并做出总结. 提出一种新的数据流预测方法, 将输入的数据流构造成有向图数据结构, 并随着新数据流的到来对有向图进行维护, 从中提取概率矩阵, 利用马尔可夫模型进行计算, 得出数据流下一时刻可能的到来值. 最后通过实验验证其可行性. 结论表明, 该算法的精度和预测成功率都较高.

关键词: 数据流预测; 有向图构造; 马尔可夫模型

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)05-0040-05

Prediction Algorithm Based on Digraphs Constructing over Data Stream

YOU Xin, YIN Gui-sheng

(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract: The forecasting techniques of data stream provide such an important role in the applications of financial portfolio management and network monitoring. The forecasting techniques of data stream in recent years are analyzed and made conclusion. A new forecasting method of data stream is proposed, which constructs the input data stream to digraphs structure. When new data stream comes, the digraphs is maintained. Then the probability matrix that describes the data stream approximately can be received, and the possible value in the next time of a data stream can be forecasted by the use of Markov models. Finally, the feasibility is verified through experiment. The conclusion indicates that the ratio of the algorithm's precision and success forecasting is high.

Key words: data stream prediction; digraphs constructing; Markov models

1 引言

数据流是指一系列数据以时间为基准组成的序列, 具有连续性、无限性等特征. 近年来人们越发关注数据流技术的应用, 并将其渗透到股票、网络监控、气象分析等领域中, 通过对数据流建立模型进行预测分析. 数据流的预测模型通常具有如下特点: (1) 模型处理的数据不是存储在硬盘中的数据, 而是由连续数据组成的流序列^[1]. (2) 数据流中的数据的先后顺序只与时间相关, 并且数据到来时只有一次处理机会. (3) 数据流中的大多数数据处理后应立即从硬盘中移除, 除非有特殊需要, 否则不能被再次提取处理(如有必要再次提取, 代价很高). 根据这些特

点, 设计的数据流预测技术需要高效、低代价^[2].

现有的数据流预测技术^[3]已取得一定的研究成果. 回归分析算法能够通过引入与待预测因子相关的其他变量提高预测精度; 在小样本的数据预测中效率很高. 缺点是在该方法中不考虑该值会随时间发生变化的因素; 另外在小样本预测中因缺乏足够的训练样本而导致预测精度偏低, 而在大样本的数据预测中, 该方法却又很难获得准确的拟合曲线或者得到该曲线的计算代价会成倍增加. 指数平滑法简单、经济、直观和容易理解. 计算简单, 快速, 所需的数据存储量等资源较少. 它的缺点是没有正式的建模算法, 使用代价较高; 只适用于变化不大的平稳时间序列; 当时间序列发生突变时, 预测结果不太理

想,而且也不适合较长时间序列的预测;无法引入与待预测因子相关的变量,不能通过这种方式提高预测精度. ARMA 模型适合用于短期预测,当样本数量很大时,得到的参数估计值也很接近. 神经网络算法通过大量样本数据建立模型,并且在数据到来时不断调整,因此预测精度较高,但计算代价较大,因此不适合实时性高的情况.

基于上述数据流预测的特点,文中提出了基于有向图构造的数据流预测算法,其基本思想是对数据流上的数据做聚类^[4],使需记录的数据集由无限变为有限,同时以时间为轴,将相邻聚类点的变化关系表示在数据流有向图中,通过计算聚类点状态转移关系的概率研究其规律,得到一个可以近似描述数据流的概率矩阵. 最后,通过应用马尔可夫模型预测数据流在下一步时刻可能的到来值.

2 基于有向图构造的预测模型

2.1 预测模型总体设计思路

该预测模型的总体思想是当输入数据到来时,通过滑动窗口模型^[5]选取需要处理的数据,并随着新的数据到来,对窗口内的数据进行更新^[6];然后对数据流应用 CluStream 算法做聚类^[7],同时构造有向图数据结构;当新数据到来时,利用有向图构造算法对其进行维护;根据从实时构造及维护的有向图中提取概率矩阵,利用马尔可夫模型进行计算,得出数据流下一时刻可能的到来值.

通过以上分析,得到该预测模型如图 1 所示.

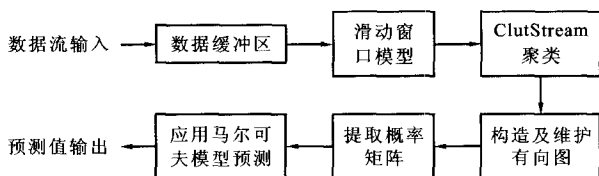


图1 预测模型

由该模型图例可知,整个预测过程有两个过程,有向图构造和应用马尔可夫模型的数据预测.

2.2 相关名词解释

数据流集合: $X = \{x_1, x_2, \dots, x_p, \dots\}$, 为了减小数据的处理代价,采用长度为 $|w|$ 的滑动窗口模型进行处理.

节点:对数据流上各点进行聚类后的簇. 记为 C_i, C_j 由四元组 (id, value, account, num) 构成,其中 id 代表该节点的编号, value 代表该节点的近似取值(在整个有向图中具有唯一性), account 表示数据流上值为 value 的聚类点个数, num 表示该节点内

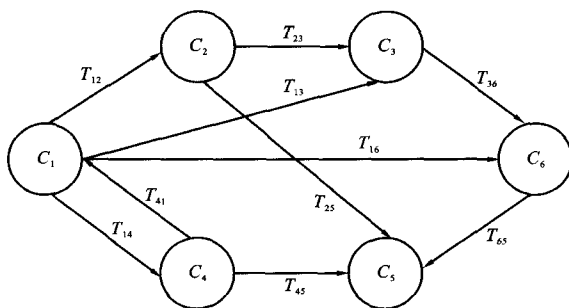
所含数据的个数.

趋势线:节点 C_i 到 C_j 的连线记为趋势线 T_{ij} , 由二元组 (id, count) 构成,其中 id 代表该趋势线的编号, count 代表发生在由节点的 C_i 到 C_j 的变化次数.

数据流有向图:由节点和趋势线构成,由该图即可近似描述一个无限数据流的变化状态,大大节省了存储和计算代价.

2.3 数据模型设计

首先选用经典的 CluStream 算法对数据流做聚类,对于每一个聚类点取一个近似值代表,使一个连续变化的曲线分割成折线,再进一步转化为直方图. 这一过程实际上是将连续的数据流分割成了多个不连续的数据块,且每个数据块大小可能不同,如果将每个不同数值、不同大小的数据块作为有向图中的一个节点,节点数目过大,因此设定 h 值,当数据块的值与某个节点的值相同,且数据块所含数据个数与节点的 num 值的差在 $[0, h]$ 之间,则该数据块在数据流有向图中可由该节点表示(具体构造算法见第 3 节). 由此,可构造数据流有向图如图 2 所示.



$$\text{到概率矩阵 } \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix}.$$

将此概率模型代入马尔可夫模型方程即可对下一时刻数据进行预测。

该算法的空间复杂度为 $O(m^2)$, 根据最坏情况下计算的时间复杂度为 $O(|\log_2^{n+1}| \times m^2)$, 由于计算乘法的所需时间非常小, 因此数据预测可以在数据流的下一到达时刻前完成。

3 有向图构造与维护算法

由于马尔可夫预测模型非常经典, 因此文中重点介绍如何构造和维护预测模型的数据结构——有向图。有向图构造与维护模型如图 3 所示。

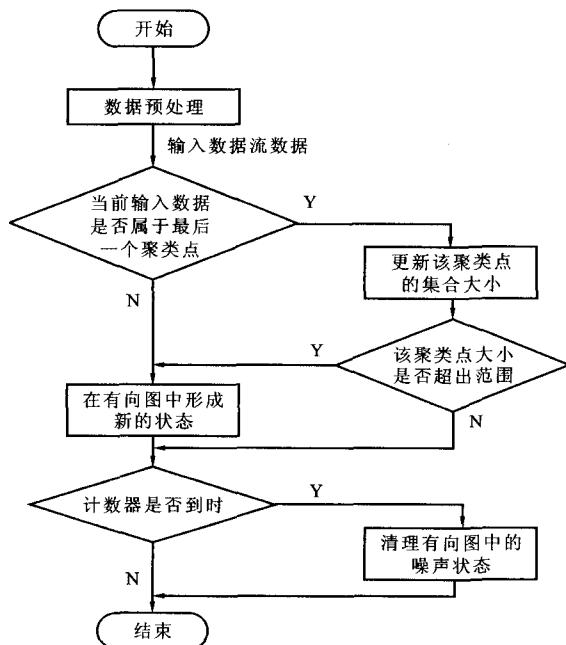


图 3 有向图构造与维护模型

3.1 输入数据处理

当没有对数据流进行确切定义时, 输入数据的值域范围可以是 $(-\infty, +\infty)$, 因此需要对输入数据进行预处理, 可以提高处理速度, 同时方便进行有向图的构造及维护。文中采取的预处理方法是归一化处理, 需要注意的是, 当得到预测值时, 需要进行数据的反归一化处理。

3.2 有向图的初始化

应用模型前使用一定数量的样本数据构造初始数据流有向图, 其过程包括两步: 聚类, 对输入历史数据流选取滑动窗口, 对选定窗口上的数据流做聚类操作, 实现对数据流的基本处理; 有向图初始化,

将各聚类点的变化关系表示在数据流有向图中。

3.3 有向图的构造与维护

应用如下算法构造与维护有向图:

(1) 查询

假设新到来的数据为 d , 根据窗口内的聚类结果首先得到 d 所属的节点 C_i , 然后在有向图中查找值为 value 的节点 C_j , 根据 C_i 的 account 和 num 值判断 C_i 是否属于 C_j , 具体查询节点方法如下:

//定义 h 为 num 的可容忍差量范围

```

boolean query(){
    for( $n=0; n < \text{有向图中节点个数}; n++$ ){
        if( $C_i.\text{value} == C_n.\text{value}$  并且
             $|C_i.\text{num} - C_n.\text{num}| < h$ ){
            return true;
        }
        return false;
    }
}
  
```

(2) 新增

当新的数据到来时, 对滑动窗口中的数据做聚类, 如果新来的数据不属于最后一个节点 C_i , 则在数据流有向图中创建一个新的节点 C_{i+1} . 具体的新增节点方法如下:

void add(C_i) { //新建节点 C_i

$C_i.\text{id} = \max(C_{id}) + 1;$

$C_i.\text{account} == 1;$

$C_i.\text{value} = \text{该数据块的数据平均值};$

$C_i.\text{num} = \text{该数据块的数据个数};$

//查找数据流中 C_i 的前一节点(利用 query 函数),

设为 C_s , 增加趋势线 Tis

$Tis.\text{id} = \max(Tid) + 1;$

$Tis.\text{count} == 1;$

}

(3) 更新

当新的数据到来时, 对滑动窗口中的数据做聚类, 如果新来的数据属于最后一个聚类点, 则更新该聚类点所属节点 C_i 的 account 属性; 如果该聚类点超出有向图聚类点的大小范围, 则在有向图中形成新状态。如发现当前聚类点和前一聚类点的变化关系符合节点到节点的趋势线, 则更新该条趋势线的 count 属性。具体的更新节点方法如下:

void update(C_i) {

$C_i.\text{account}++;$

//查找数据流中 C_i 的前一节点(利用 query 函数),

设为 C_s , 更新趋势线 Tis

$Tis.\text{count}++;$

(4) 去噪

由于噪声或其他不确定因素的影响,数据流上可能出现噪点^[8],这些噪点的节点在数据流有向图中出现次数非常少;另外一些历史久远的节点及其涉及到的趋势线可能在很长一段时间内没有再出现.对于以上两类节点我们称之为无效节点,这类节点对于数据流的预测没有正确指导意义,反而会造成预测的误判断.需要定期找出这样的节点,并执行删除操作.判断噪声节点的算法如下:

```
totalcount =  $\sum_n^{节点数} C_n$ .account // 定义变量
void isNnode() {
    for(n=0; n<有向图中节点个数; n++) {
        if(Cn.num < totalcount/10 × 节点数
            或者
            Cn的趋势线最近更新时间 > 50) {
            delete(该节点);
        }
    }
}
```

(5) 删除

有向图的维护可以进行删除操作,删除对象包括噪声节点或某些被强制要求删除的节点.删除时,除删掉该节点外,还应删除与其相关的趋势线.删除节点方法如下:

```
void delete(Cn) {
    while(Tni.next()) { // 删除节点相关趋势线
        删除趋势线 Tni;
    }
    删除 Cn 节点本身;
}
```

4 实验及性能分析

为了验证算法的预测效果,本节中将对算法的性能进行分析并通过实验数据进行对比.文中选用较常用的实数流对算法进行验证,数据来源分别是:(1)股票数据在线网提供的上海证券交易所数据,使用其中2004年6月至2009年6月的大盘周线数据(2)中国气象科学数据共享服务网提供的“中国农业基本气象资料月值数据集”,使用其中1991年5月至2006年2月中国地面观测站的月报基本气象段数据.

性能指标选取平均绝对误差(MAD)、平均相对误差(MRD)、数据预测成功的比例(Ratio)三项作为分析标准.

(1) h 值对数据流预测精度的影响.如图4所示,数据1随着聚类粒度的增大,预测精度先增大后减小,而数据2随着聚类粒度的增大,预测精度慢慢

减小.第一组实验数据相对于第二组实验数据变化较剧烈,规律性较弱,而第二组数据有内在规律,在一定时间内数据值的变化较小,且变化连续,所以第二组数据预测精度较高.对于第一组数据,当 $h=10$ 时,预测精度最高.对于第二组数据,当聚类粒度较小时,数据流有向图的节点数较多,每个节点的趋势线也较多,能更好的描述数据的变化趋势,因此当聚类粒度较小时,聚类精度更好.

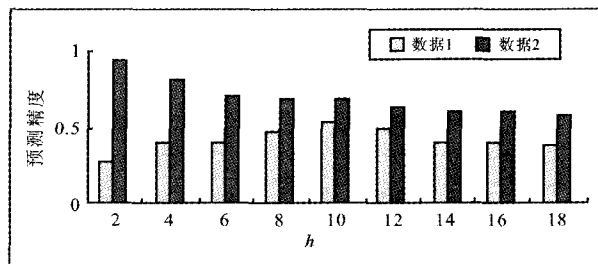


图4 h 值对数据流预测精度的影响

(2) 预测时长对数据流预测精度的影响.图5为在第二组实验数据上进行的预测曲线和实际数据曲线对比图的缩略显示,由图可见,随着预测时长的增大,预测的误差也呈增大趋势.针对此现象进行分析,得出原因:进行时间段数据预测时,需要对先到来的数据进行预测,同时使用该预测结果去预测后续数据可能的变化态势.因此当预测时长大于1时(即预测的时间点大于1个),由于误差量逐渐累积,时间段预测的误差将会越来越大,且增长迅速.

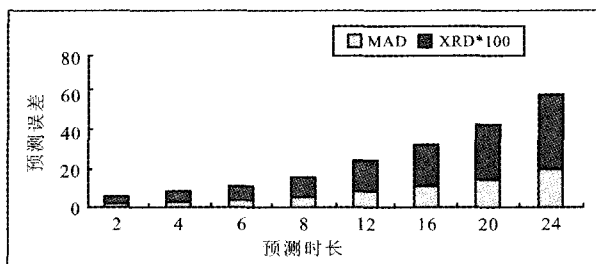


图5 预测时长对数据流预测精度的影响

(3) 最后将本算法与改进的EPM算法分别在大量样本数据和少量样本数据情况下的预测结果进行对比分析,选用第二组数据作为数据输入.改进的EPM算法采用的预测方法是回归模型,我们在前文中提到过,当样本数据较小时这种方法效率较高,但是预测的精度低;当样本数据较大时,预测结果与实际数据流的拟合度较低,而且计算成本大.文中算法采用的预测核心思想是基于概率统计模型的,因此当样本数据较小时,预测精度较差,而当样本数据逐渐增多时,预测精度与实际数据的拟合度会越来越好,且该方法的计算代价不随样本数据增多而增大.

如图 6、图 7 所示,在样本数据较小的情况下,文中方法不如 EPM 算法精度高;而在样本数据较大的情况下,文中方法具有较高的优势。

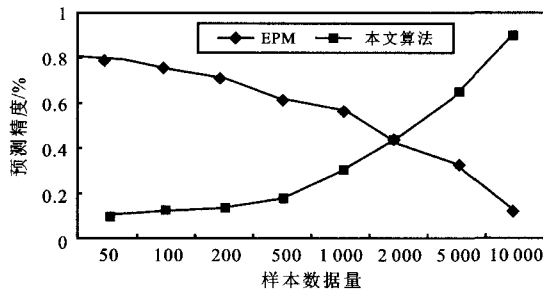


图 6 样本数据量对预测精度的影响

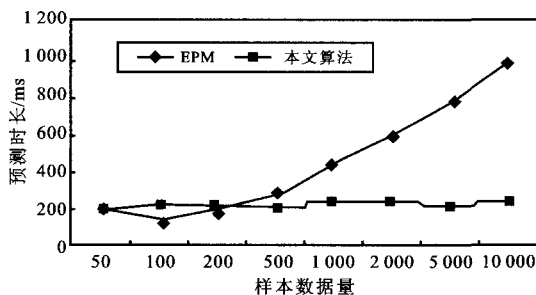


图 7 样本数据量对预测时长的影响

5 结束语

实验结果表明,该算法能够高效地记录并描述数据流的变化态势,并对数据流在下一时刻的到来数据进行预测.它适用于类似金融、气象等具有连续性,且前后变化幅度可控的情况,及要求预测效率较高,但预测时长较短的情况;特别在样本数据量较大的情况下,文中方法具有较高的优势.相较其他同类算法而言,该算法的精度和预测成功率都较高.但该算法仍存在局限性和不足,如对于变化幅度较大且不连续的数据流预测效率较低;不适用于需要预测出未来较长时间内情况变化的数据;另外,有向图构造时,采用的是经典数据流聚类方法,该方法本身仍存在不足,可进一步改进以提高基于有向图构造的

数据流预测算法效率.

参考文献:

- [1] Domingos P, Hulten G. Mining high-speed data streams[J]. Knowledge Discovery and Data Mining, 2000;71-80.
- [2] Dong Guozhu, Han Jiawei, Laks V S, et al. Online mining of changes from data streams: research problems and preliminary results[C]// Workshop on Management and Processing of Data Streams. In cooperation with ACM SIGMOD/PODS and FCRC. San Diego, CA, USA: ACM, 2003.
- [3] 刘耀宗,王湛. 数据流的预测与分类研究[J]. 计算机科学, 2007, 34(11): 170-173.
- [4] 刘海峰,刘守生,姚泽清,等. 一种基于改 K-means 聚类的文本特征选择模型[J]. 微电子学与计算机, 2009 (6): 29-31.
- [5] 周黔,吴铁军. 一种动态数据流的实时趋势分析算法[J]. 控制与决策, 2008(10): 1182-1185.
- [6] Yu Meng, Margaret H Dunham. Efficient mining of emerging events in a dynamic spatiotemporal environment[J]. Lecture Notes in Computer Science, 2006 (3918): 750-754.
- [7] Aggarwal J, Han J, Yu P S, et al. A framework for clustering evolving data streams[C]// Proceedings of the 29th International Conference on Very Large Data Bases (VLDB) Berlin, Germany: VLDB, 2003: 81-92.
- [8] 王树广. 分布式数据流上的连续异常检测[J]. 微电子学与计算机, 2008, 50(9): 158-160.

作者简介:

由欣女, (1983-), 硕士研究生. 研究方向为计算机软件与理论数据库与知识库.

印桂生男, (1964-), 教授, 博士生导师. 研究方向为数据库、虚拟现实、信息安全.

(上接第 39 页)

- [6] Yen ShaWei, Hu MingChih, Chen ChihLung, et al. A 0.92mm² 23.4mW fully-compliant CTC decoder for WiMAX 802.16e application [C]// Custom Integrated Circuits Conference, CICC '09. San Jose, CA: IEEE, 2009: 191-194.
- [7] 吕闻. 基于 FPGA 的高效 CTC 译码器的设计和实现[D]. 西安: 西北工业大学, 2007.

作者简介:

曹旭 (1985-), 硕士研究生. 研究方向为专用集成电路设计与系统应用.

郝学飞 (1976-), 副研究员. 研究方向为无线通信算法, 无线通信芯片设计.

胡国荣 (1966-), 研究员. 研究方向为多媒体通信, 信道信源芯片设计.