

基于 I-GARCH 的不确定时间序列概率分布推算

汤其婕, 王 珂

(南京航空航天大学 计算机科学与技术学院 江苏 南京 211106)

摘要: 处理不确定数据存储问题的常用方法是使用概率数据库的方法,但是已有的概率数据库生成方法是针对已知概率分布的数据集。不确定时间序列在每一时刻的概率分布规律随着时间的变化而变化,无法使用统一的概率推导方法进行计算,因此已有的概率数据库生成方法不再适用。为解决该问题,依托已有的 ARMA 模型和 GARCH 模型,提出推导不确定时间序列概率分布的推算模型。同时,为了进一步增强该模型的容错性,提出了相应的错误值过滤算法。实验结果表明,该模型能够有效地根据不确定时间序列的发展规律,动态地进行调整计算,得出不确定时间序列的概率分布;同时,容错算法能够很好地探测到数据集中的错误数据,进行数据的清洗与替换,体现出良好的容错性与一般通用性。

关键词: 不确定时间序列; 概率分布推算; ARMA 模型; GARCH 模型; 错误值过滤

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2018)12-0023-06

doi: 10.3969/j.issn.1673-629X.2018.12.005

Probability Distribution Estimation of Uncertain Time Series Based on I-GARCH

TANG Qi-jie, WANG Yu

(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
Nanjing 211106, China)

Abstract: One of the most effective ways to deal with uncertain time series is to employ probabilistic database. But existing methods of generating probabilistic database are typically based on the assumption that the probabilistic distribution is already known. The law of probability distributions of uncertain time series varies with time and cannot be calculated by a uniform method, so the existing methods of generating probabilistic database are no longer applicable. To solve the problem based on the existing ARMA model and GARCH model, we propose a prediction model for deriving the probability distribution of uncertain time series. In addition, in order to enhance the fault tolerance of the model, we propose a corresponding erroneous value filtering algorithm. Experiment shows that the model can effectively adjust and calculate the probability distribution of uncertain time series, which is in line with the development law of the origin uncertain time series. Furthermore, the erroneous value filtering algorithm can detect and find the erroneous values well, and then wash and replace the data with inferred correct values, which shows great fault tolerance and commonality.

Key words: uncertain time series; probability distribution estimation; ARMA model; GARCH model; erroneous value filtering

0 引言

时间序列(time series)是一种典型的高维数据类型,其在传感器网络、位置定位服务(location based service, LBS)、环境监测、医疗检测、物联网等众多领域应用广泛^[1-2]。但是,受数据采集设备的缺陷或者人为因素的影响,采集得到的数据在一定范围内存在偏差。将这类型的数据定义为不确定时间序列(uncertain time series)。而针对不确定时间序列的有效存储,到目前为止仍没有良好的解决方案。

一种处理不确定数据最有效的方案是概率方法。近年来许多专家和学者提出了一系列的方法用于解决不确定数据的管理和查询问题^[3-9]。这些方法有一个共同特征,即假定用于进行查询的概率数据是已知的,可以直接获取到。但是,现实情况并非如此。不确定时间序列的概率是由推导出这些概率的概率分布函数决定的,这些概率分布函数以时间为坐标不断发生变化。简而言之,不确定时间序列的概率值随时间不断变化,无法得到其固定值,因此无法使用已有的概率数

收稿日期: 2017-11-23

修回日期: 2018-04-18

网络出版时间: 2018-06-29

基金项目: 国家自然科学基金(61772269)

作者简介: 汤其婕(1994-),女,硕士研究生,研究方向为数据与知识工程;王 珂(1994-),男,硕士研究生,研究方向为数据与知识工程。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20180629.1700.004.html>

数据库生成方法直接对其进行存储处理。因此,如何“创造”概率数据仍是未解决的问题,也是文中主要研究的问题。

针对从已知的时间序列推导得到时间序列的概率分布问题,文中主要完成的工作分为两部分。一是依托已有的各种数学模型,结合已有的动态密度指标的概念,提出 ARMA-GARCH 动态密度指标模型,并对其推算原理进行了详细的分析与介绍;二是针对 GARCH 模型无法高效处理含错数据的弊端,提出改进的 I-GARCH 模型。该模型在处理含错数据集时能体现出良好的容错性,更符合一般的不确定时间序列数据的采集规律。最后通过实验进行验证。

1 相关工作

1.1 动态密度指标

时间序列由于依赖时间的变化,通常呈现出很大的不确定性,因此为不确定时间序列数据创建概率数据库的最大挑战之一是处理不断更新的概率分布。如图 1 所示,该图为一天的气温随时间的变化曲线。

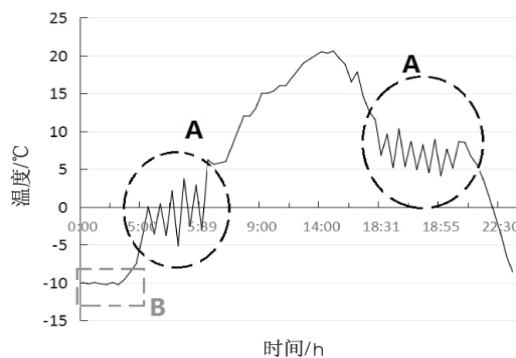


图 1 气温变化曲线

在临近日出和日落两个时间点,温度的变化十分明显,如 A 区域所示,但是在夜间的时候,整体的温度变化幅度不大,如 B 区域所示。两处的概率分布规律明显不一致。如果用相同的概率分布基来表示两处的概率分布,显然不科学。因此,应该随着时间的变化动态地更新用来表示概率分布的概率分布基,使其符合当前数据的变化趋势。由此,引入了动态密度指标的概念。

动态密度指标^[10]依托多种数学模型,可以从一条给定的时间序列中动态地推算出随着时间变化的概率分布。然后,由这些动态密度指标推算得到的概率分布就可以用来创建概率数据库,完成数据的存储工作。已有的动态密度指标介绍如下。

1.2 已有动态密度指标

(1) 统一阈值指标。

Cheng 等^[11]提出了一个通用的不确定数据的查

询估约框架。它的主要思想是将原始数据进行建模,将建模得到的数据范围作为对应时间点数据的波动范围。同时,该范围也是该时间点所对应的真正值的所在范围。然后在计算出的波动范围中进行查询操作,代替直接在原始值上进行查询。

统一阈值指标(uniform thresholding metric, UT)的思想是上述思想的一种扩展,即通过推导得到对应时间点的“预期真实值”,然后以该“预期真实值”代表该时间点的原始真实数值,表示该点的概率分布。“预期真实值”的定义如下。

定义 1(预期真实值):给定一个概率密度方程 $p_i(R_i)$, 预期真实值 \hat{r}_i 是 R_i 的期望值,定义为 $E(R_i)$ 。

统一阈值指标将用户定义的阈值 u 作为统一分布的界限,其中心为推导得到的真实值。用户定义的 u 为波动范围,真实值 \hat{r}_i 和它对应的原始值 R_i 的差值保证在 u 以内。

期望真实值的推算采用了自回归移动平均模型(autoregressive moving average, ARMA)^[12-13]。例如,给定一条时间序列 $S = \langle r_1, r_2, \dots, r_t \rangle$ 和一个滑动窗口 S_{t-1}^H , 得到 $r_i = r_t + a_i$, $t - H \leq i \leq t - 1$, 并且 a_i 服从零均值正态分布,方差为 σ_a^2 。给定一个 ARMA(p, q) 模型,推算预期真实值 \hat{r}_i 为:

$$\hat{r}_i = \varphi_0 + \sum_{j=1}^p \varphi_j r_{t-j} + \sum_{j=1}^q \theta_j a_{t-j} \quad (1)$$

其中, (p, q) 是非负整数,定义了模型的顺序; $\varphi_1, \varphi_2, \dots, \varphi_p$ 是自回归参数; $\theta_1, \theta_2, \dots, \theta_q$ 是移动平均参数; φ_0 是一个常量; $t > \max(p, q)$ 。

(2) 可变阈值指标。

可变阈值指标(variable thresholding metric, VT)与统一阈值指标的区别在于两点。一是可变阈值指标基于高斯分布,而统一阈值指标基于均匀分布;二是可变阈值指标不需要用户提供阈值范围,而是给窗口 S_{t-1}^H 计算样本方差 s_t^2 , 然后建立高斯分布模型^[14]。

给定窗口 S_{t-1}^H , 动态阈值指标推算得到一个正态分布为:

$$p_i(R_i = r_i) = \frac{1}{\sqrt{2\pi s_i^2}} e^{-\frac{(r_i - \hat{r}_i)^2}{2s_i^2}} \quad (2)$$

其中, \hat{r}_i 是由 ARMA 模型推导得出的预期的真实值。

2 GARCH 指标

由上述可知,统一阈值指标中 u 是一个固定值,这与实际情况不相符,因为在真实世界中,每个时间点的波动范围通常不是一个统一值,而是随着时间的变化不断发生改变。由图 1 可以看出,区域 A 数据波动明显,而区域 B 数据波动较为平缓。区域 A 和区域 B

数据的波动规律不一致,在进行数据的表示时不能使用统一的概率密度方程笼统代替。通过进一步研究发现,在进行一个概率密度函数推算时,底层的描述模型加入均值和时变方差能够很好地提高数据描述的精度,由此提出了 GARCH 密度指标的概念。

GARCH 将 $p_i(R_i)$ 建模为一个高斯概率密度方程 $N(\hat{r}_i, \hat{\sigma}_i^2)$, 该指标认定时间序列不仅仅表现出与均值 \hat{r}_i 有关,同时也与方差 $\hat{\sigma}_i^2$ 有关。 \hat{r}_i 的计算由 GARCH 模型得出, $\hat{\sigma}_i^2$ 的计算则通过 ARMA 模型推算得出。

2.1 GARCH 模型

GARCH^[15] 模型 (generalized autoregressive conditional heteroskedasticity) 很好地体现出一条时间序列中的数据波动特征。给定窗口 S_{t-1}^H , ARMA 模型 $r_i = \hat{r}_i + a_i$, 其中 $t-H \leq i \leq t-1$, 定义条件方差 σ_i^2 :

$$\sigma_i^2 = E((r_i - \hat{r}_i)^2 | F_{i-1}) \quad \sigma_i^2 = E(a_i^2 | F_{i-1}) \quad (3)$$

其中, F 是已知的 $i-1$ 时间前的所有的信息; $E(a_i^2 | F_{i-1})$ 是 a_i 的方差。则 GARCH(m, s) 模型将式 3 建模为一个关于 a_i^2 的线性方程:

$$a_i = \sigma_i \varepsilon_i \quad \sigma_i^2 = \alpha_0 + \sum_{j=1}^m \alpha_j a_{i-j}^2 + \sum_{j=1}^s \beta_j \sigma_{i-j}^2 \quad (4)$$

其中, ε_i 表示独立分布的随机变量序列; (m, s) 表示描述模型顺序的参数; $\alpha_0 \geq 0, \alpha_j \geq 0, \beta_j \geq 0$, $\sum_{j=1}^{\max(m, s)} (\alpha_j + \beta_j) < 1$, i 的范围在 $t-H + \max(m, s)$ 和 $t-1$ 之间。

GARCH 模型反映出 a_i 的波动会造成整个模型的波动。与可变阈值指标中的方差 s_i^2 不同, σ_i^2 是在 \hat{r}_i 之后计算。在很多实际应用中, GARCH 模型通常为 GARCH(1, 1)。为了推导出时间变化的幅度, 采用的 GARCH(m, s) 模型如下:

$$\hat{\sigma}_i^2 = \alpha_0 + \sum_{j=1}^m \alpha_j a_{i-j}^2 + \sum_{j=1}^s \beta_j \sigma_{i-j}^2 \quad (5)$$

\hat{r}_i 的推导使用到了 ARMA 模型, 因此提出了 ARMA-GARCH, $\hat{\sigma}_i^2$ 的计算由 GARCH 模型推导得出。 \hat{r}_i 和 $\hat{\sigma}_i^2$ 的具体推导过程如下:

算法 1: 使用 ARMA-GARCH 推导 \hat{r}_i 和 $\hat{\sigma}_i^2$

输入: ARMA 模型参数 (p, q)、滑动窗口 S_{t-1}^H 和比例因子 k

输出: $\hat{r}_i, \hat{\sigma}_i^2$ 和边界参数 u_b, l_b

1. 推算模型 ARMA(p, q) 得到 a_i , 其中 $t-H + \max(p, q) \leq i \leq t-1$

2. 根据 a_i 推算模型 GARCH(1, 1)

3. 根据 ARMA(p, q) 推导出 \hat{r}_i , 根据 GARCH(1, 1) 推导出 $\hat{\sigma}_i^2$

4. $u_b \leftarrow \hat{r}_i + k\hat{\sigma}_i, l_b \leftarrow \hat{r}_i - k\hat{\sigma}_i$

5. return $\hat{r}_i, \hat{\sigma}_i^2, \mu_b, l_b$

算法 1 给出了 \hat{r}_i 和 $\hat{\sigma}_i^2$ 的详细推算过程。 \hat{r}_i 使用 ARMA 模型推导得出, $\hat{\sigma}_i^2$ 使用 GARCH 模型推导得出 (步骤 3)。其中 $k \geq 0$ 是决定 u_b, l_b 的因子, 假如设定 $k=3$, 则 r_i 分布在 u_b, l_b 之间的概率十分高, 约为 0.997。步骤 1 和步骤 2 的时间复杂度分别为 $(H \cdot \max(p, q))$ 和 $(H \cdot \max(m, s))$ 。与 H 相比 (p, q) 很小, 因此整体的估算效率很高。

2.2 加强的 GARCH 模型 I-GARCH

在实际中, 时间序列通常存在噪声点或者错误值, 例如传感器错误、网络断开等。上述提出的 GARCH 模型只适用于处理数据精确的不确定时间序列, 对于包含错误数据的时间序列没有很好的性能。为了解决这一问题, 提出了一种加强的 GARCH 密度指标 I-GARCH(improved GARCH)。

为了更好地说明 GARCH 模型针对数据轨迹发生变化时的处理, 针对时间序列 $S = \langle r_1, r_2, \dots, r_t \rangle$, 在所有的窗口 S_{t-1}^H 上运行 ARMA-GARCH 算法, 设置 $k=3$ 。结果显示如图 2(a) 所示。A 表示推导得到的界限范围, 从图中可知, A 的范围远远超出正常范围, 即用 ARMA-GARCH 处理出现严重错误, 该方法明显不适用。并且从图中可知, 在 $t=96$ 时, 出现错误数据, 此时, ARMA-GARCH 在之后的几个时间点出现极大的波动。这主要是因为 GARCH 表达式中 (式 4) 出现了平方式, 这会增加错误值对整个模型的影响。为了避免出现这个情况, 提出了改进的 GARCH 模型, 该模型能够检测出一条原始数据中包含的错误数据, 将其从原始数据中删除, 用推导得出的值代替原有值。

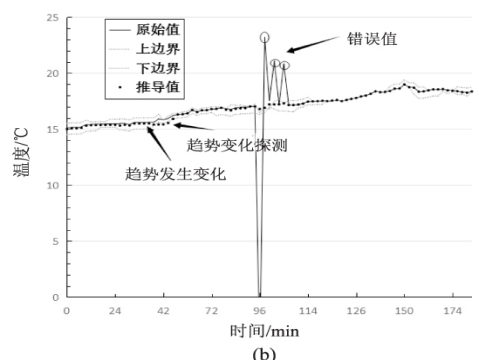
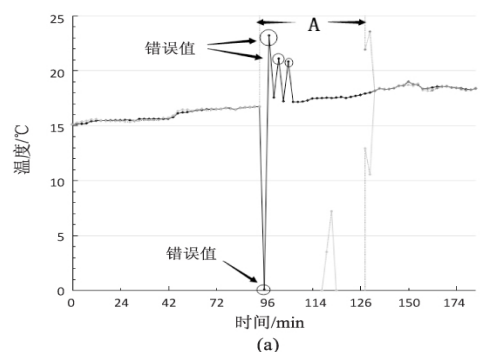


图 2 ARMA-GARCH 和 I-GARCH 举例说明

假设 $S = \langle r_1, r_2, \dots, r_{t_n} \rangle$ 为一条包含了错误值的时间序列。由算法 1 计算得到在 $t > H$ 的 ARMA-GARCH 过程。设置 $k = 3$, 保证 r_t 超出 u_b, l_b 界限的概率很低。当出现 r_t 超出 u_b, l_b 界限时, 将其标注为错误值, 并且用相应的推算值 \hat{r}_t 代替。同时, 记录出现连续错误值的个数, 如果该数目超出了预定义的常 E_{\max} , 则认定序列的发展趋势出现了高速的变动。例如, 在日出和日落阶段, 温度的变化波动较大, 就会出现连续的“错误值”, 这时, 就需要重新调整模型, 使其符合新的趋势。

3 I-GARCH 动态指标的改进

尽管在现实中, 一条时间序列连续出现错误值的可能性很小, 但是为了确保数据的精确性, 提出了一种新的方法, 用来过滤 I-GARCH 模型中的错误值, 称为错误值过滤算法 EVF (erroneous value filtering)。

算法的输入为包含错误值的时间序列 $V = [v_1, v_2, \dots, v_k]$ 以及阈值参数 DT_{\max} 和 E_{\max} 。具体的实现步骤如下:

(1) 计算记录了一条时间序列 V 中, 两两相邻的数据之间的差值;

(2) 遍历差值集合, 根据 DT_{\max} 判断该差值是否在允许范围内, 如果小于阈值参数, 默认该数值为正确值; 如果差值大于阈值, 则继续遍历;

(3) 如果连续出现差值超过阈值的情况, 记录出现的次数, 如果该次数大于 E_{\max} , 则认为这些连续的点并非错误值, 而是时间序列的走势发生了明显的变化, 原始数值不作变动, 继续向下遍历;

(4) 反之, 当记录的次数在阈值范围内, 则说明该点为异常点。找到该点在序列中的位置, 将其删除。并通过线性插值的方法计算新的值代替原有错误值。

算法 2: EVF

输入: 包含错误值的时间序列 V , 差值阈值 DT_{\max} , 连续错误个数阈值 E_{\max}

输出: 干净值序列 V

```
1. ArrayList<Double> differList = new ArrayList<>();
2. int differ = 0; int count = 0;
3. for( int i = 1; i < k; i++) {
4. differ = abs( vi - vi-1 );
5. differList.add( differ );
6. }
7. ArrayList<Integer> posList = new ArrayList<>();
8. int i = 0;
9. while( i < k ) {
10. int count = 0;
11. while( differList.get( i ) < DTmax && i < k ) { i++; }
12. while( differList.get( i ) > DTmax && i < k ) { count++; i++;
```

```
++;
```

```
13. if( count < Emax ) { countList.add( i ); } //为异常点, 记录其位置
```

```
14. }
```

```
15. for( int i = 0; i < posList.size(); i++) {
```

```
16. Vi+1 为错误值将其删除;
```

```
17. 使用( vi + vi+2 ) / 2 线性插值代替 Vi+1;
```

```
18. }
```

4 实验

4.1 实验目的

实验目的主要有两个: 验证提出的动态密度指标 ARMA-GARCH 对于真实数据集有良好的准确性与高效性; 比较 ARMA-GARCH 与 I-GARCH, 验证添加了错误过滤的 I-GARCH 模型对处理包含错误数据的数据集的优越性。

4.2 实验数据

实验数据取自两个真实的数据集。一个是 Temperature Dataset, 该数据集记录了 20 天内传感器网络监测得到的气温变化的所有数据, 约 21 000 条样本数据。另一个数据集为 GPS Dataset。这个数据集包括从导航系统记录的 192 辆车的 GPS 日志。每一个日志元组包含时间和 $x - y$ 数值, 本实验只取用其中的 x 数值。该数据集包含约 10 000 条数据。两个数据集的详细情况如表 1 所示。

表 1 实验数据说明

属性数据集	Temperature Dataset	GPS Dataset
检测参数	温度	GPS 定位
样本数目	21 085	10 473
传感器精确度	$\pm 0.3^\circ\text{C}$	$\pm 10\text{ m}$
样本时间间隔	2 min	1 ~ 2 s

4.3 实验方法

(1) 动态密度指标的衡量。

假设 $p_t(R_t)$ 是 t 时刻推算得到的概率密度方程, 衡量该概率密度方程最直接的方法是将其与真实的概率分布方程 $p_t(R_t)$ 进行对比, 但是一般情况下无法获得 $p_t(R_t)$ 。因此, 文中采用了概率积分变换的方法间接估算 $p_t(R_t)$ 。给定一个随机变量 X 及其概率密度方程 $f(X)$, 通过 $Y = \int_{-\infty}^x f(X=u) du$ 将其转化为一个均匀分布的随机变量 Y 。因此, 关于 $p_t(R_t)$ 的概率积分变换为 $z_t = \int_{-\infty}^{r_t} p_t(R_t = u) du$ 。

设 $p_1(R_1), p_2(R_2), \dots, p_t(R_t)$ 是用动态密度指标推导得到的概率分布序列 z_1, z_2, \dots, z_t 为相应的概率积分变换值。则只有当 $p_t(R_t)$ 等于真正的密度分布 $p_t(R_t)$ 时, z_1, z_2, \dots, z_t 才会均匀分布在 $(0, 1)$ 之间。实

验使用直方图近似法验证 z_1, z_2, \dots, z_t 的累计分布方程, 判断其是否为均匀分布, 将其累计方程定义为 $O_z(z)$, 同时定义在 $(0, 1)$ 上均匀分布的标准累计方程为 $U_z(z)$ 。定义 $O_z(z)$ 和 $U_z(z)$ 之间的差距为密度距离, 表达式如下:

$$d\{U_z(z), O_z(z)\} = \sqrt{\sum_{x=0}^1 [U_z(x) - O_z(x)]^2} \quad (6)$$

密度距离可以量化地测量观察值分布 z_1, z_2, \dots, z_t 和它们的预期分布之间的差距, 因此可以作为衡量动态密度指标的标准。

(2) 实验过程。

第一部分: 动态密度指标的比较。

将提出的 ARMA-GARCH 与统一阈值和可变阈值进行比较。所有的评估都在两个数据集上进行。使用密度距离作为衡量各个动态密度指标质量的标准。同时, 也比较了各动态密度指标的运行效率, 以运行时间作为衡量的标准。

第二部分: I-GARCH 和 ARMA-GARCH 的比较, 实验在 Temperature Dataset 上进行验证。为了比较两个指标对于处理数据的精确性, 在原有数据中插入人工合成的错误数值, 即随机地在原始数据中插入数值远高于或低于正常范围数据的数值。以捕获错误值的数目和运行时间作为衡量两个指标优劣的标准。

4.4 实验结果与分析

(1) 第一部分的实验结果。

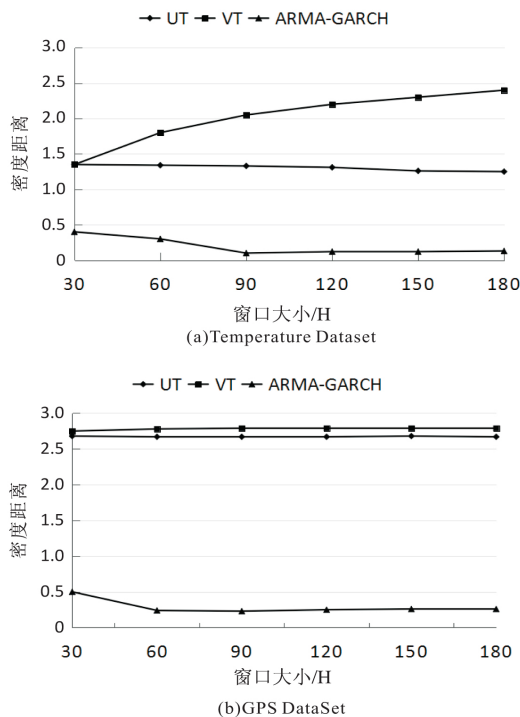


图 3 动态密度指标比较

图 3 显示了随着窗口尺寸 (H) 的增大, 各种动态密度指标在两个数据集上的密度距离的比较。从图中

可以明显看出, MARA-GARCH 优于原始的动态密度指标。

图 4 显示了执行一次密度推算迭代所需的平均时间。由图中可以看出, 虽然 ARMA-GARCH 的执行时间总体上超出原始动态密度指标, 但是差距并不明显, 大约在 0.2 ~ 0.4 s 左右。考虑到其在准确度和效率上的优势, ARMA-GARCH 仍是性能最好的动态密度指标。

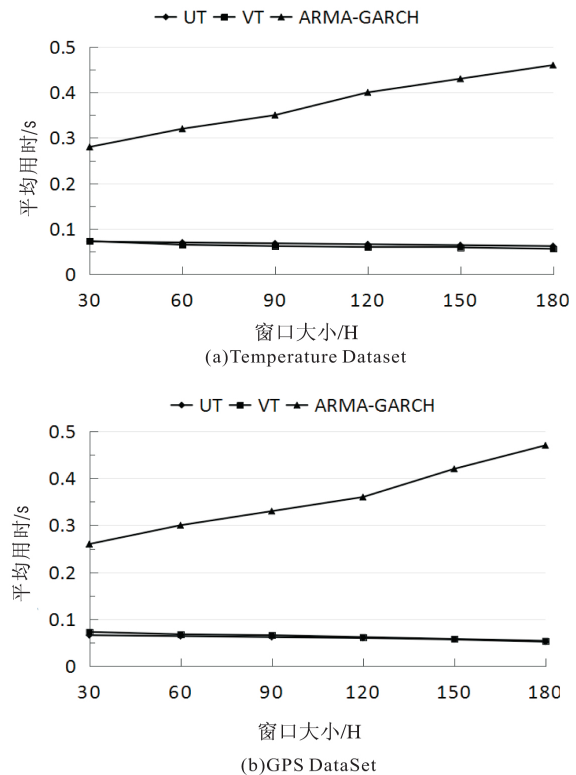
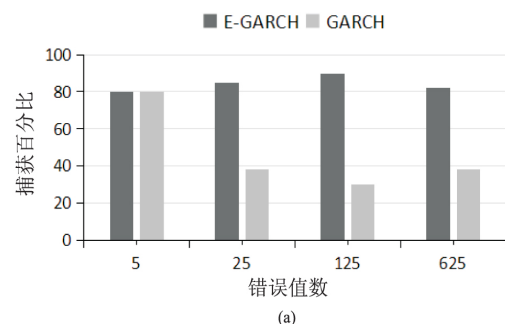


图 4 动态密度指标效率比较

(2) 第二部分的实验结果。

图 5 比较了 I-GARCH 和 ARMA-GARCH 检测得到的错误值的比例。从图 5 (a) 很明显可以看出, I-GARCH 在检测和清除错误值方面的性能是 ARMA-GARCH 的两倍。同时, 从图 5 (b) 可以看出, 与 ARMA-GARCH 相比, I-GARCH 所需的时间消耗更小。这是因为, 当错误值发生在窗口 S_{t-1}^H 中时, ARMA 模型需要消耗更多的时间。这个额外的时间抵消了 I-GARCH 模型清除错误值的时间。



(a)

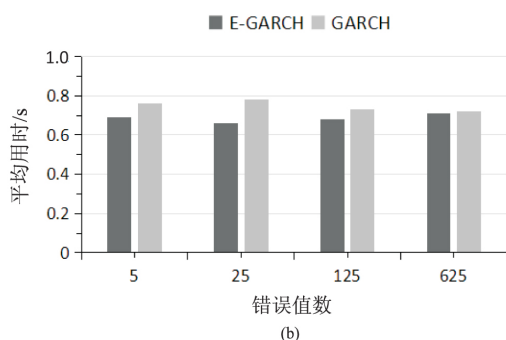


图 5 I-GARCH 和 GARCH 的比较

综上,文中提出的 ARMA-GARCH 模型及 I-GARCH 模型与已有的统一阈值指标(UT)以及可变阈值指标(VT)相比具有很大的优势,可以准确地推算出不确定时间序列的概率密度分布,在准确度和时间消耗上优势明显;同时,优化了 I-GARCH 指标,提出的算法 EVF 可以很好地检测出不确定时间序列中的错误值,进行错误值的清洗与替换,具有良好的容错性和一般通用性。

5 结束语

不确定时间序列的概率分布随着时间的变化而不断改变,无法使用已有的概率数据库生成方法直接对其进行数据库生成操作。因此在进行数据的存储之前,需要对原始数据进行有效的概率分布推导工作,得到不确定时间序列数据随着时间变化的一般分布规律。文中依托已有的 ARMA 模型和 GARCH 模型,提出推导不确定时间序列概率分布的 ARMA-GARCH 模型以及 I-GARCH 模型,并且在此基础上进行进一步的改进,提出能有效过滤错误值的算法 EVF。实验结果表明,ARMA-GARCH 模型和 I-GARCH 模型能有效地根据时间序列的发展规律推导得出正确的概率分布。同时,针对包含错误数据的数据集,EVF 算法体现出高效的错误排查功能,具有良好的容错性和一般通用性。下一步的研究工作是利用推导得出的概率分布生成不确定时间序列的概率数据库。

参考文献:

[1] 王小平,罗军,沈昌祥.无线传感器网络定位理论和算法[J].计算机研究与发展,2011,48(3):353-363.

[2] CHENG R, KALASHNIKOV D V, PRABHAKAR S. Querying imprecise data in moving object environments[J]. IEEE Transactions on Knowledge & Data Engineering, 2004, 16(9): 1112-1127.

[3] 周傲英,金澈清,王国仁,等.不确定性数据管理技术研究综述[J].计算机学报,2009,32(1):1-16.

[4] 蒋涛,高云君,张彬,等.不确定数据查询处理[J].电子学报,2013,41(5):966-976.

[5] 崔斌,卢阳.基于不确定数据的查询处理综述[J].计算机应用,2008,28(11):2729-2731.

[6] 江彤,金宗安,谢东.概率数据库的聚集查询[J].计算机工程,2010,36(11):42-44.

[7] ANDRITSOS P, FUXMAN A, MILLER R J. Clean answers over dirty databases: a probabilistic approach[C]//Proceedings of the 22nd international conference on data engineering. [s. l.]: IEEE, 2006: 30.

[8] RE C, LETCHNER J, BALAZINKSA M, et al. Event queries on correlated probabilistic streams[C]//Proceedings of the 2008 ACM SIGMOD international conference on management of data. Vancouver, Canada: ACM, 2008: 715-728.

[9] OLTEANU D, HUANG J, KOCH C. SPROUT: lazy vs. eager query plans for tuple-independent probabilistic databases[C]//International conference on data engineering. Shanghai, China: IEEE, 2009: 640-651.

[10] SATHE S, JEUNG H, ABERER K. Creating probabilistic databases from imprecise time-series data[C]//IEEE 27th international conference on data engineering. Hannover, Germany: IEEE, 2011: 327-338.

[11] CHENG R, XIA Y, PRABHAKAR S, et al. Efficient indexing methods for probabilistic threshold queries over uncertain data[C]//13th international conference on very large data bases. [s. l.]: [s. n.], 2004: 876-887.

[12] AKAIKE H. Maximum likelihood identification of Gaussian autoregressive moving average models[J]. Biometrika, 1973, 60(2): 255-265.

[13] 邹柏贤,刘强.基于 ARMA 模型的网络流量预测[J].计算机研究与发展,2002,39(12):1645-1652.

[14] 季铎,王智超,蔡东风,等.基于高斯分布的簇间距离计算方法[J].中文信息学报,2008,22(3):50-55.

[15] SHUMWAY R H, STOFFER D S. Time series analysis and its applications[M]. [s. l.]: Springer, 2009.