

不同粒度时间序列相似性度量

邵校莎莎¹,郝文宁¹,靳大卫¹,王莹²

(1.解放军理工大学工程兵工程学院,南京 210007; 2.武警海南省总队一支队,海口 570000)

(shao_sxss@sina.com)

摘要: 现有的时间序列的相似性度量大多基于欧氏距离,并不适用于不同粒度时间序列的相似性匹配,无法直接对其相似性进行有效的度量,为此,提出一种基于对应差值比样本的相似性度量,用于不同粒度时间序列的相似性匹配。首先对不同时间粒度的时序数据进行阐述,并定义了对应差值比样本与相似度计算方法;接着提出基于它们的相似性匹配算法;最后实验证明,该度量能够有效地度量不同粒度时间序列数据的相似性。

关键词: 时间序列;相似性度量;时间粒度

中图分类号: TP301.6 **文献标志码:** A

Similarity measurement of time-series data with different granularities

SHAO Xiao-shasha¹, HAO Wen-ning¹, JIN Da-wei¹, WANG Ying²

(1. Engineering Institute of Corps of Engineers, PLA University of Science and Technology, Nanjing Jiangsu 210007, China;

2. The first Detachment of CAPF Hainan Province Corps, Haikou Hainan 570000, China)

Abstract: Most of the existing similarity measurement, based on Euclidean distance, cannot be applied directly and effectively to similarity matching of the time-series with different granularities. This paper proposed a new similarity measure based on the sample of the corresponding D-value. It firstly expounded the definition of the time-series with different granularities, and defined the sample of the corresponding D-value; secondly it put forward the similarity matching algorithm; finally, the experimental results prove that the algorithm can effectively measure the similarity of time-series with multiple granularities.

Key words: time-series; similarity matching; time granularity

0 引言

时序数据是指按时间顺序取得的一系列观测值。从经济到工程技术,从天文到地理和气象,几乎各个领域都存在时间序列^[1]。时间序列的数据挖掘主要包括时序数据的相似性挖掘、离群挖掘和规则挖掘等,其中相似性挖掘是目前时序数据的研究热点^[2]。时间序列的相似性度量是衡量两个时间序列的相似程度的方法;它是时间序列分类、聚类、异常发现等诸多数据挖掘问题的基础,也是时间序列数据挖掘的核心问题之一^[3]。现实世界中,大量存在的时序数据,由于需求、存储或采集过程中的观测等诸多因素,时间间隔不可能是统一的长度。因而,在时间序列相似性搜索中,一个重要问题是时间的描述和划分,时序数据的时间粒度就是衡量时间序列数据的时间单位,通常用时态类型来表示时间粒度。常用的时间粒度有秒、分、时、日、月和年等。在一些实际应用中,也经常用不同时间粒度去描述同一类问题^[4]。

根据时间序列本身的特点,度量它们的相似性会有不同的方法。由于不同粒度时序数据本身所特有的性质,目前已有的时间序列相似性度量(大多基于欧氏距离),不能直接有效地对其进行相似性比较,本文将于第1章对此进行阐述说明。

本文主要研究了不同时间粒度的时序数据的相似性匹配问题,而关注的重点在于不同粒度时序数据动态变化趋势的

相似性匹配。首先阐述了相关定义及经典相似性度量,并提出一种新的基于对应差值样本的相似性度量;它主要关注动态变化趋势的相似性,有效规避了基于距离的相似性度量在解决不同粒度时间序列相似性匹配问题上的不足。接着提出不同时间粒度时间序列相似性匹配算法;然后做实验对所提算法进行验证。

1 相关工作

目前已有的相似性度量主要有欧氏距离(L_p)^[5-6]、动态时间弯曲距离(Dynamic Time Warping, DTW)^[7-8]、最长公共子串(Longest Common Subsequence, LCSS)^[9]、实序列编辑距离(Edit Distance on Real Sequence, EDR)^[10]等。而 L_p 与DTW是两种经典的相似性度量,应用也最为成熟广泛。

L_p 是把时间序列看成相同维的两个点,计算它们的距离作为两个时序数据的相似度,并设定一个阈值来评判是否相似。欧氏距离虽然简单实用,但它有两个内在特点:1)要求两个序列等长而且两个序列的值必须是一一对应;2)在判断过程中完全忽略时间因素,仅把时间序列看成一个多维点。

这两个特点限制了 L_p 在不同粒度时序数据的相似性问题中的应用:首先,不同粒度的时间序列一般不会等长,其序列值也无法一一对应。其次,由于其完全忽略了时间因素,一般能较好地运用于两个时间跨度相等,时间粒度相同的时序数据,因为时间因素在此类时间序列相似性匹配中并不产生

收稿日期:2011-06-29;修回日期:2011-08-05。

作者简介:邵校莎莎(1987-),女,江西鹰潭人,硕士研究生,主要研究方向:作战效能评估、军用数据工程;郝文宁(1971-),男,山西运城人,副教授,博士,主要研究方向:军用数据工程、海量高维数据规约、作战效能评估;靳大卫(1979-),男,河北保定人,讲师,主要研究方向:军事运筹学、军事数据工程、作战效能评估;王莹(1985-),女,湖南永州人,主要研究方向:军需管理工程。

影响,可以忽略。但由于不同粒度的时间序列的值点是基于不同的时间单位的,如果忽略其中时间因素,直接将这两个时序数据视作多维点并计算它们之间距离作为相似度,必然扩大了其差异性,而不能准确地表现它们的相似度。例如,一个公司的月销售数据与季销售数据。一个季度的销售量是三个月销售量的累计和,那么直接用月销售值与季销售值计算其欧氏距离,总体上必然扩大了其值的差异,而无法关注其动态趋势的相似,产生不合理结果。

而动态时间弯曲距离(DTW)虽然允许一对多点的距离比较,但其仍然是将时间序列纯粹看成一个多维点,忽略其中的时间因素。而大多其他简单的相似性度量是欧氏距离及其其他的变种,基于共同的准则,更多地关注于值的总体相似性,而对于不同粒度时间序列这种,因为时间单位不同而单个值差大,总体趋势却相似的数据,势必得不到合理结果。针对这一问题,本文提出了一种新的基于差值比例的相似性度量,它能有效避免距离度量的这一缺陷,为直接对不同粒度时序数据相似性匹配提供有效的度量。

2 问题描述

本文着重讨论的是不同时间粒度的数值型一元时间序列数据。

由于两个时间序列的时间粒度不同,进行点对点的比较是不合理更是不可行的。因而进行比较前,首先进行序列对应点划分,再计算其对应差值比样本。

2.1 相关定义

易知,粒度相对大的序列,其相邻值点的时间跨度就大。例如,假设有时序数据 A 和 B , A 的时间粒度为季, B 的时间粒度为月,则 A 相邻值点的时间跨度为一个季度,而 B 相邻值点的时间跨度仅为一个月。因而,如果将 A 、 B 视作二维折线段画入一个坐标系内(X 轴用统一的时间单位),则折线 B 可被 A 的时间点划分成几段,每段由 B 中的几个值点组成,对应一个 A 中的点,这样就把 A 与 B 中的值点分别对应起来,进而计算对应的差值比样本。

定义 1 给定两个不同时间粒度的时间序列数据 Q_a, Q_b :

$Q_a = \{(q_{a1}, t_{a1}), (q_{a2}, t_{a2}), \dots, (q_{am}, t_{am})\}$, 时间粒度为 tg_a , Q_a 的观测点个数为 m , 总时间跨度为 T_a ;

$Q_b = \{(q_{b1}, t_{b1}), (q_{b2}, t_{b2}), \dots, (q_{bn}, t_{bn})\}$, 时间粒度为 tg_b , Q_b 的观测点个数为 n , 总时间跨度为 T_b ;

其中 q_{ai}, q_{bj} 分别表示 Q_a 在时刻 t_{ai} 及 Q_b 在时刻 t_{bj} 的观测值, $T_a = T_b$ 。假设时间粒度单位 $tg_a < tg_b$ 且 $Ntg_a = 1tg_b$, 则可用 Q_b 的点将 Q_a 中的值点划分成 n 份:

令 $k = [(t_{ai} - t_{a1})/N] + 1$, 则 Q_a 中的第 i 个点被划分入第 k 段, 即对应于 Q_b 中的第 k 个点 Q_{bk} 。则令 $b_i = q_{bk} - q_{ai}$ ($i = 1, 2, \dots, m$); 计算下列 m 个值: $c_i = \frac{b_i}{\text{aver}(b_i)}$ ($i = 1, 2, \dots, m$) 并依次设为 c_1, c_2, \dots, c_m , 称 $\Sigma = \{c_1, c_2, \dots, c_m\}$ 为两个时间序列的对应差值比样本。

为了利用两个时间序列的对应比值样本判断它们是否相似, 文献[11]中提出了一种基于折线图形相似的相似性度量, 根据两曲线的几何相似来判断时间序列的相似性。借鉴此思想, 提出以下相似性判断推论。

推论 已知两个不同时间粒度的时间序列数据的对应比值组为 $B = \{b_i; i = 1, 2, \dots, m\}$, 时间序列完全相似的充分必要条件是 b_i 等于一个常数 B 。因而对应差值比样本 $\Sigma =$

$\{c_1, c_2, \dots, c_m\}$ 中 $c_i = 1$, 那么时间序列相似的充分必要条件为: $c_i = 1$ 。

因为现实应用中, 时序数据的相似往往仅要求检验统计量或趋势的相似性, 因而不需要 $c_i = 1$ 严格成立, 只需 c_i 变化幅度在一个较小的控制范围内, 则判定两时间序列相似。本文假设对应差值比样本服从均值为 1 的正态分布, 设定一个置信区间, 若 c_i 均落于这个区间则视作该等式成立, 并用差值比样本中落于这一区间的 c_i 在总数 m 中的比例作为两个时间序列的相似度度量。

定义 2 若两个不同粒度的时间序列数据的对应差值比样本为 $\Sigma = \{c_1, c_2, \dots, c_m\}$, Σ 的均值为 1, 均方差为 σ 。如果有 c_i 落在相应的 95% 的置信区间时, 即有: $c_i \in [1 - 1.96\sigma, 1 + 1.96\sigma]$ 那么就称时间序列 Q_a 中的第 i 个点与时间序列 Q_b 相似, 记作 $Q_{ai} \sim Q_b$ 。

定义 3 给定两个不同时间粒度的时间序列数据 Q_a, Q_b , 其中 Q_a, Q_b 的时间粒度分别为 tg_a, tg_b 且 $tg_b < tg_a$ 。若 Q_a 中有 k 个点都有 $Q_{ai} \sim Q_b$, 则称 Q_a 与 Q_b 的相似度为 k/m (m 为 Q_b 的观测点数)。

2.2 不同时间粒度时间序列相似性匹配算法

不同时间粒度时间序列算法共包括两个子算法: 对应差值比样本计算算法和相似性判定算法。待相似性匹配的两个时序数据首先由对应差值比样本计算算法处理, 计算出不同时间粒度的时序数据的对应差值比样本; 然后把处理后的结果输入到相似性判定算法中, 利用对应差值比样本的正态分布特性, 计算样本中值落入设定置信区间的点的比率, 即两时间序列的相似度。最后将此值与已定阈值进行比较判断两个时序数据是否相似。

2.2.1 对应差值比样本计算算法

由于待相似性匹配的两个时间序列的时间粒度不一样, 则相邻点的时间跨度与序列长度均不一样, 无法直接将两序列的点对应, 并利用原始时间序列数据计算两者距离; 因而必须首先运用合适的对应策略将两序列点的进行对应处理, 再通过计算每组对应点的差值与差值的算术均值, 确定对应差值比样本。

对应差值比样本计算算法伪代码如下:

输入: Sequence $Q_a[2][m]$ 表示时序数据 Q_a , m 个点; Sequence $Q_b[2][n]$ 表示时序数据 Q_b , n 个点, 其中 $Q_{a/b}[0][i]$ 存放观测值, $Q_{a/b}[1][i]$ 存放 $Q_{a/b}[0][i]$ 观测点在该组时序数据中的按时间先后排列的序数; H 表示 Q_a 与 Q_b 时间粒度换算值。

输出: Sequence C 表示 Q_a 与 Q_b 的对应差值比样本。

- 1) i 从 0 开始循环到 $n-1$ 时, 执行步骤 2);
- 2) $k = Q_b[1][i]/H, Q_b[1][i] = k, Q_b[0][i] = Q_a[0][k] - Q_b[0][k]$;
- 3) 计算 $Q_b[0][i]$ 的均值 $\text{aver}g_b$
- 4) i 从 1 开始循环到 n , 在新的时序数据中计算出对应差值比样本 $c_i = Q_b[0][i]/\text{aver}g_b$;
- 5) 输出 Q_a 与 Q_b 的差值比样本 C 。

2.2.2 相似性判定算法

不同粒度时间序列 Q_a 与 Q_b 经对应差值比样本计算算法处理后, 得到对应差值比样本 $C = \{c_1, c_2, \dots, c_n\}$ 。如果两个时间序列相似, C 中值的变化应在一个较小幅度内。由于对应差值比样本具有正态分布性, 而 C 的均值 μ 为 1, 均方差为 σ , 因而规定如果 c_k 均落在相应的 95% 的置信区间时, 即有: $c_k \in [\mu - 1.96\sigma, \mu + 1.96\sigma]$, 则判定 Q_a 与 Q_b 相似。由于数据中可能存在噪声, 为使算法具有更好的鲁棒性, 在算法实现过

程中加设一个阈值,只要 C 中令 $c_k \in [\mu - 1.96\sigma, \mu + 1.96\sigma]$ 成立的点个数大于一定比率(所设定的阈值),则认为 Q_a 与 Q_b 相似。这就允许时间序列中存在少量噪声点,而不至影响最终结果。

相似性判定算法伪代码如下:

输入: Q_a 与 Q_b 的对应差值比样本 C ;差值比样本 C 中值的个数 H ;
判定是否相似阈值 B 。
输出:相似度量 sim 。
1) 计算 C 的均方差 σ ;
2) 赋初值: $k = 1, p = 0$
3) 判断 $k \leq H$ 是否成立,成立则执行4),不成立转5);
4) 先判断 $c_k \geq \mu - 1.96\sigma$ 且 $c_k \leq \mu + 1.96\sigma$ 是否成立,成立则 $p++$,再执行 $k++$,转3);
5) 计算 $sim = p/H$,输出 sim ;
6) 判断 $sim \geq B$ 是否成立,成立则输出“similar”,不成立则输出“not similar”。

3 实验结果与分析

3.1 实验数据

本文实验数据来自 <http://www.bundesbank.de/>^[12]网站,它是证券市场板块下的 Time series WU0053: Gross sales of domestic debt securities at nominal value/Total。两个时间序列数据都选自1985年~1989年,时序数据 Q_a 的取样周期为每月一次,共60个数据项;时序数据 Q_b 每一个季度测值一次,共20个数据项。两时间序列数据的折线图,如图1所示。

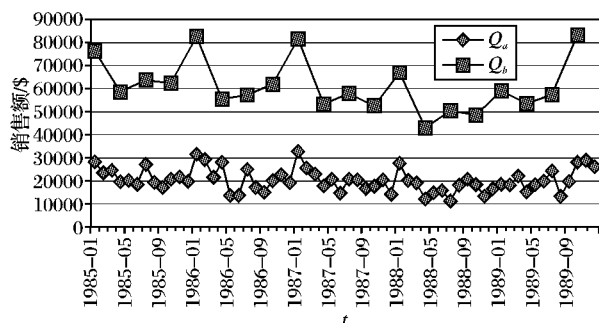


图1 时序数据 Q_a 、 Q_b 折线图

图1中,时序数据 Q_a 与 Q_b 横坐标 t 取值范围为1985—01—1989—12,间隔为一个月。显然,采用传统的基于度量的相似性匹配方法,很可能判断这两个不同时间粒度时序数据为不相似。然而,实际上, Q_a 与 Q_b 是同一 debt securities 的不同时间粒度的两组观测值,本质上是同一个事物,应是完全相似。也就是说,传统相似性度量在直接运用于两个不同时间粒度的时序数据时往往会产生不合理的结果,不具有完全可用性。

3.2 实验结果

本文采用基于对应比值样本的新的相似性度量对实验数据进行相似性判定。首先将数据输入对应差值比样本计算算法,得到差值比样本 C ,本文仅通过列出差值比样本 C 主要统计量,对 C 进行必要描述:总样本数为60,有效样本数60,全距为0.8276,极小值为0.6629,极大值为1.4905,均值1.0,标准差0.2009。

比值样本 C 的频率直方图如图2所示,标准流差为0.201, $N = 60$ 。图2中频率柱状图为 C 取值在区间 $[0.6, 1.6]$ 频数的分布,线条是均值为1的高斯分布。从图中不难看出, C 的取值绝大多数落在1附近,即绝大多数的数据项取值相当接近;只有少数取值偏离1较远。

为进一步验证,将差值比样本 C 及相似性阈值 B 输入相

似性检验算法,本文取相似度阈值为0.9。经算法处理得相似度 $sim = 57/60 = 0.95$ (即有57点落于95%置信区间内),大于相似度阈值0.9,故这两个时间序列是相似的。由于实验所选数据是同一 debt securities 的不同时间粒度的两组观测值,本质上是同一个事物,进一步证实了算法的可信性。

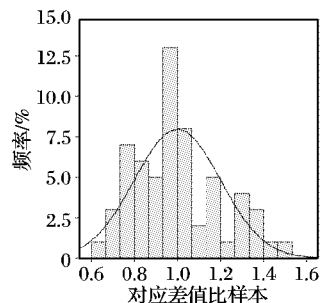


图2 差值比样本 C 频率直方图

4 结语

本文针对不同时间粒度的时间序列数据相似性匹配问题,提出了一种较为合理的解决思路:首先对不同时间粒度时间序列的点进行对应处理;然后计算对应差值比样本;最后利用差值比样本的正态分布特性,计算落于预先设定的置信区间内点的比率,作为两序列相似度量。实验证明,该算法计算简单,具有很好的鲁棒性,并支持增量计算,能判断两个不同时间粒度时序数据的相似性。

参考文献:

- [1] 阎继伟. 时间序列的数据挖掘研究[D]. 上海: 上海交通大学, 2006.
- [2] CHAN K P, FU A W. Efficient time series matching by wavelets [C]// Proceedings of the 15th IEEE International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 1999: 126—133.
- [3] 刘懿, 鲍德沛, 杨泽红, 等. 新型时间序列相似性度量方法研究[J]. 计算机应用研究, 2007, 24(5): 112—114.
- [4] AGRAWAL R, LIN K I, SAWHNEY H S, et al. Fast similarity search in the presence of noise, scaling, and translation in time-series databases[C]// Proceedings of the 21st International Conference on Very Large Databases. San Francisco: Morgan Kaufmann Publishers, 1995: 490—501.
- [5] KEOGH E. Similarity search in massive time series databases[D]. Gainesville: University of California, Department of Computer Science, 2001.
- [6] KEOGH E, RATANAMAHATANA C A. Exact indexing of dynamic time warping[J]. Knowledge and Information Systems, 2008, 7(3): 358—386.
- [7] 陈当阳, 贾素玲, 王惠文, 等. 时态数据的趋势分析及其子序列匹配算法研究[J]. 计算机研究与发展, 2007, 44(3): 516—520.
- [8] VLACHOS M, KOLLIOS G, GUNOPOULOS D. Discovering similar multidimensional trajectories[C]// Proceedings of the 18th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2002: 673—684.
- [9] CHEN L, OZSU M T, ORIA V. Robust and fast similarity search for moving object trajectories[C]// Proceedings of the ACM SIGMOD Conference. New York: ACM, 2005: 491—502.
- [10] 周勇, 林甸. 时间序列相似性的图形相似方法研究[J]. 统计与决策: 理论版, 2007(10): 28—30.
- [11] 刘宁. 时态数据库多时间粒度问题的研究[D]. 哈尔滨: 哈尔滨理工大学, 2005.
- [12] Time series[EB/OL]. [2011—02—20]. http://www.bundesbank.de/statistik/statistik_zeitreihen_en.php?lang=en&open=&func=row&tr=WU0053.