

基于加权 IMF 对时间序列相似匹配^{*}

孙汝儒, 肖迪

(南京工业大学 自动化与电气工程学院, 南京 211800)

摘要: 经验模态分解(EMD)算法非常适合非稳定序列信号、非线性序列信号以及复杂信号的分解, 具有很高的噪声比。序列信号经过 EMD 分解为本征模函数(IMF)以及残差序列, 所分解出来的 IMF 包含了原序列信号不同时间尺度的局部特征信号, 是整个原序列的“去杂”反映。针对 IMF 所包含的不同尺度的特征这一特性, 给出用 EMD 分解原始序列信号, 提取其全部有限个本征模函数和残差序列, 根据不同的 IMF 所包含原序列的特征信息量的大小引入信息权重 w , 然后通过欧氏距离对各个序列不同 IMF 序列进行相似匹配判定, 最后通过综合各个 IMF 所占权重综合判定时间序列的相似匹配。实验结果表明, 基于 IMF 对时间序列相似匹配和直接对原时间序列进行匹配, 前者首先对时间序列进行分解, 去掉其噪声等干扰, 提取出 IMF 间接进行加权匹配, 提高了时间序列的模式匹配精度, 证明了该方法的有效性。

关键词: 经验模态分解算法; 本征模函数; 本征模函数加权; 相似模式匹配

中图分类号: TP301.6 **文献标志码:** A **文章编号:** 1001-3695(2013)12-3664-03

doi: 10.3969/j.issn.1001-3695.2013.12.037

Time series similarity matching based on weighted IMF

SUN Ru-ru, XIAO Di

(College of Automation & Electrical Engineering, Nanjing University of Technology, Nanjing 211800, China)

Abstract: Empirical mode decomposition (EMD) algorithm is very suitable for non-stable decomposition of the sequence of signals, nonlinear sequence signal, and complex signals with high noise ratio. Sequence signal after EMD decomposition the costs intrinsic mode functions (IMF) and the residual series decomposition IMF contains the local features of the different time scales of the original sequence signal is the original sequence to the complex reflects. Aiming at the IMF contained different scales characteristic of this feature, this paper gave EMD decomposition of the original sequence signal, extract all of its limited number of intrinsic mode functions and residual series, the amount of information contained in the original sequence characteristics depending on the IMF the size of the introduction of the right to information, and then used the Euclidean distance to match the different IMF sequences for each sequence similarity, at last it comprehensively judged the time series similarity matching of each IMF's weight. The experimental results show that, time series similarity matching and directly matching to the original time series based on IMF, the former method firstly decomposed the time series to remove the noise and other interference, extracted IMF weighted matching indirectly improve the time series pattern matching accuracy, to prove the effectiveness of this method.

Key words: EMD algorithm; intrinsic mode functions(IMF); IMF weighted; similar pattern matching

0 引言

时间序列^[1]是指将某种现象某一种统计指标在不同时间上的各个数值, 依据时间顺序排列形成的一种序列。时间序列分析被广泛地应用在金融市场、产品销售数据、工业工程、水利事业、故障诊断检测以及飞行器安全飞行等各个领域。并且随着社会的发展, 人们对已知和未知的认识要求的提高, 对时间序列分析的应用与分析俨然成为当今生活不可或缺的重要研究部分。但是如何在海量的数据库中通过相似性匹配应用分析来获取信息和知识, 这是一个具有重要意义的理论和实际应用课题。这对未知的探求、科学决策、异常行为检测以及认识事物的发展具有重要的指导意义。

对于时间序列相似匹配, 学者都作了深入研究, 如离散傅

里叶变换法(DFT)^[2]将时间序列转换成一系列多维空间的点, 针对点与点之间进行距离比较匹配, 它在变换过程中会忽略高频信号, 把序列信号的局部极大值、极小值变得平滑, 因此, 这样就会对序列信号形成遗漏, 许多重要信息也因此丢失, 这就造成了它的局限性。动态弯曲距离(DTW)^[3]相比于欧氏距离是可以允许信号序列在时间轴上对模式进行变换伸缩, 用来查询模式 R 和给定的模式 Q 相似匹配。还有其他许多相似匹配的方法, 但是大部分时间序列相似匹配的方法都是基于时间序列本身作出处理, 并没有看到时间序列背后所包含的信息。

本文针对时间序列信号中包含的信息进行筛选, 利用经验模态分解^[4]方法, 对原始时间序列信号进行特征提取, 即提取出本征模函数(IMF)。原始序列经过分解可以分解出有限个 IMF, 每个 IMF 序列上所包含的信息含量各不相同, 对其加权

收稿日期: 2013-04-06; **修回日期:** 2013-05-12 **基金项目:** 国家自然科学基金资助项目(51205185); 江苏省博士后科研计划基金资助项目(1001027B)

作者简介: 孙汝儒(1987-), 男, 江苏连云港人, 硕士研究生, 主要研究方向为金融风险预测、系统建模(27131342428@qq.com); 肖迪(1975-), 女, 河北大城人, 副教授, 博士, 主要研究方向为模式识别、数据挖掘、系统优化等。

处理,然后通过欧氏距离求取 IMF 的相似序列,最后找出时间序列相似匹配。该方法与直接对时间序列匹配相比,匹配精度明显提高,对复杂序列的匹配具有很好的效果。

1 EMD 和欧几里德距离

1.1 经验模态分解(EMD)

经验模态分解^[4]是一种新的应用于非线性与非平稳时间序列信号的分解方法。本质上讲就是产生一系列不同尺度的本征模函数和一个残余分量。对序列进行平稳化处理,即进入“筛”过程,需要满足两个条件,成为 IMF 分量需要的限制条件:

a) 对于整个时间序列来讲,极值的个数与穿过零点的个数需要相同或者最多相差一个。

b) 在信号任意一点局部最大值包络线与局部极值构成的下包络线的平均值为零。

满足上述特征的信号就是一个 IMF 分量。本征模函数记为 $c_i(t)$, 计算过程如下^[5]:

a) 初始化:令 $r_0(t) = x(t)$, $i = 1$ 。

b) 求出第 i 个本征模函数:

(a) 初始化: $h_0(t) = r_{i-1}(t)$, $k = 1$;

(b) 提取 $h_{k-1}(t)$ 的局部极小值和极大值点;

(c) 由 $h_{k-1}(t)$ 的极值点构成两个序列;

(d) 采用三次样条插值算法分别对(b)的两个序列进行插值,求取 $h_{k-1}(t)$ 的上、下包络线;

(e) 计算 $h_{k-1}(t)$ 上下包络线的平均值,记序列 $m_{k-1}(t)$, 并定义 $h_k(t) = h_{k-1}(t) - m_{k-1}(t)$;

(f) 如果 $h_k(t)$ 满足本征模函数的两个条件,则定义 $c_i(t) = h_k(t)$; 反之,则令 $k = k + 1$, 跳转到(b)继续重新计算。

c) 定义: $r_i(t) = r_{i-1}(t) - c_i(t)$ 。

d) 如果 $r_i(t)$ 小于预定值,算法结束;反之, $i = i + 1$, 跳转到步骤 b)。

序列 $x(t)$ 表达式为

$$x(t) = \sum_{i=1}^n c_i(t) + r_n(t) \quad (1)$$

依据经验模态分解的思想,IMF 是逐层地提取信号在每个局部的频率分量,提取是从高到低,由高频信息到低频信息,这种“走势”反映的是信号在各个局部的差异。

1.2 欧几里德距离

本文所说的“相似”是依据距离函数来衡量的,它是衡量相似的一种依据,距离越近说明序列越相似,反之,则不相似。序列之间的距离尺度采用的每个时间序列之间的距离尺度是采用欧氏距离,定义如下:

$$\text{dist}(x, y) = \left\{ \sum_{i=0}^{n-1} (X_i - Y_i)^2 \right\}^{\frac{1}{2}} \quad (2)$$

根据 Parseval 定理^[5],本文采用经验模态分解的方法,那么在该分解的区域序列之间的欧氏距离应该比原始时间序列之间的距离小。即

$$\text{dist}(x, y) > \text{dist}(X, Y) \quad (3)$$

那么,如果 $\text{dist}(x, y) < \varepsilon$, 则必存在 $\text{dist}(X, Y) < \varepsilon$ 。这样的条件限制就保证了搜索的时间序列不会存在瑕疵的结果,不会出现满足条件却被遗忘的正确的序列集合。

为了在数据库中找到所有与原始序列相似的时间序列, EMD 分解提取的函数必须满足^[11]

$$D(E(S), E(Q)) \leq D(S, Q) \quad (4)$$

其中: $D()$ 是距离函数; $E()$ 是分解的本征模函数,也就是特征函数; S, Q 是时间序列。

本文检索序列的相似依据是:距离函数 $D()$ 是欧氏距离,把 K 线对齐序列作为特征提取函数 $E()$, 满足上述定理。即两个 K 线对齐序列的欧氏距离不会超过原始序列之间的距离,因此,两个不相似的 K 线对齐序列,它们的原始序列也一定不相似。假设有原始序列 Q , 有 m 个时间序列等待查找,则存在 K 线对齐序列 $(Q_{k_0}, Q_{k_1}, \dots, Q_{k_m})$, 对于每一个 K 线对齐序列 Q_{k_i} , 它与时间序列 Q 相似, 当且仅当 $D(Q_{k_i}, Q) < \varepsilon$, ε 是阈值,完成相似模式匹配。

2 基于 IMF 对时间序列相似匹配

2.1 本征模函数的提取

原时间序列经过 EMD 分解之后,得到本征模函数以及残差序列。图 1 就是一个典型的本征模函数。

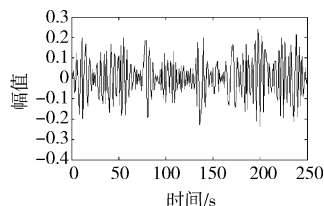


图1 本征模函数

图 1 是一个典型的本征模函数,但是原始时间序列信号分解出的往往不止一个本征模函数。根据 EMD 分解准则可知,一般对时间序列进行 EMD 分解,求本征模函数时,只有当残差序列变成单调函数时,EMD 才会停止分解进程。如图 2 所示,一个风速信号分解出的六个本征模函数以及一个残差序列。

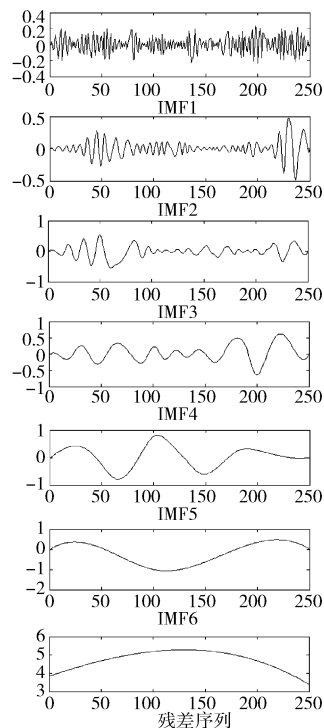


图2 风速序列经 EMD 分解后的 IMF 和残差序列

如图 2 所示,经 EMD 分解后的时间序列是由有限个本征模函数和残差序列所组成。但是这里出现一个问题,就是不同的时间序列分解出的本征模函数的个数不一定相同,当运用欧氏距离进行距离断定 IMF 的距离就会出现不相同的 IMF 匹配

困难的情形。

2.2 对本征模函数加权

如图2所示,每个时间序列信号经过EMD分解后会产生有限个本征模函数,这里并不知道不同的时间序列产生的本征模函数产生的个数,因此,需对IMF进行加权处理。

设一个时间序列经过EMD分解之后产生 n 个IMF,分别为: $(\text{imf}_1, \text{imf}_2, \dots, \text{imf}_n)$, 存在权重 w 属于 $(0,1)$ 的随机分布。由式(1)可以得出

$$x(t) = \text{imf}_1 + \text{imf}_2 + \dots + \text{imf}_n + r_n(t) \quad (5)$$

从式(5)中可以看出,原始时间序列由 n 个本征模函数和残差序列组成,这就是说原始时间序列中所包含的信息量也是其组成。这里首先假设序列所含的信息量可以进行量化。存在时间序列 $x(t)$ 含有信息量记为 $I(t)$,一个本征模函数所含有的信息量记为 I_{imf_n} ,残差序列所含信息量记为 I_m ,则存在

$$I(t) = I_{\text{imf}_1} + I_{\text{imf}_2} + \dots + I_{\text{imf}_n} + I_m \quad (6)$$

从式(6)中可知,信息量可以进行累加,这样就可以对本征模函数进行加权处理。由图2可以看出不同的本征模函数含有的信息量不等,显然,经过分解后第一次的本征模函数(IMF1所含的信息最多),所以在运用欧氏距离进行匹配时必须考虑信息量的问题。设时间序列 $x(t)$,经EMD分解后存在 n 个IMF,对其进行加权,数学表达式为

$$I_{f(x)} = \frac{1}{2} \omega \text{imf}_1 + \frac{1}{2^2} \omega \text{imf}_2 + \dots + \frac{1}{2^n} \omega \text{imf}_n + (1 - \sum_{i=1}^n \frac{1}{2^i}) r_n(t) \quad (7)$$

式(7)是本征模函数和残差序列加权后的数学表达式。

2.3 基于IMF对时间序列相似匹配

基于IMF对时间序列相似匹配^[6-10]过程如下:

- EMD分解样本时间序列和数据库序列;
- 分别提取各个时间序列的本征模函数以及趋势序列;
- 依据式(7)对各个本征模函数进行加权;
- 根据欧氏距离判断各个本征模函数的相似距离,并存储;
- 根据依据所各个本征模函数占权重综合判断IMF的相似距离;
- 找出与原序列相似匹配的时间序列;
- 匹配结束。

3 实验

实验所用数据来自国信证券的股票数据,这里选取上证指数中的450种股票从2007.5.15~2008.3.7共280天的交易数据,这里仅仅选取收盘价格作为实验数据。分别利用本文方法和利用欧氏距离直接对序列进行距离匹配的方法最后进行匹配。在匹配结果中以第111序列为例比较两种方法的准确性。本文方法对时间序列相似匹配,实验结果如图3所示。

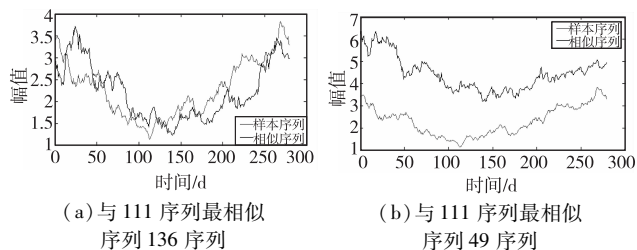


图3 对时间序列相似匹配

用时间序列直接进行相似匹配实验结果如图4所示。

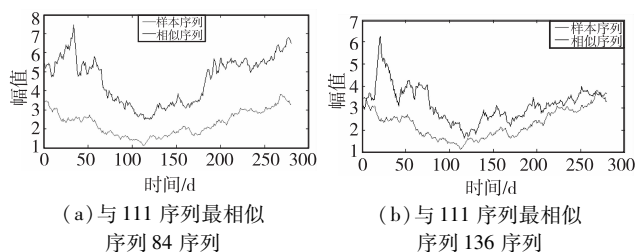


图4 时间序列直接相似匹配

从图3和4中可以清楚地看出时间序列的相似程度。图3(a)是与样本序列最相似序列,从图中可以看出,依据加权本征模函数距离的判定相似,很好地去掉了噪声之类的干扰,从原始序列所含有的特性入手,搜寻最相似的序列。而从图4中可以看出,用时间序列直接进行匹配仅仅看的是序列本身的相似程度,而不能依据序列本身所含的特性进行相似匹配搜索。图4(b)中则很好地证明了这一点,无关乎序列的特征,只是序列本身相似或是距离最近,因此导致搜索结果存在误差。

4 结束语

本文首先从时间序列的本质特性出发,用EMD算法对时间序列进行分解处理,从而提取IMF,因为IMF含有原时间序列的大部分特征信息,分解最后去除残差序列,而原时间序列的大部分特征则保留下来。由于各个IMF所包含的信息量各不相同,分解到最后IMF所含的信息量相对较小。因此,本文提出对IMF进行加权,最后通过对IMF运用欧氏距离判定法,去判定各个IMF的相似程度,然后根据各IMF所占的权重去综合判定原始时间序列的相似匹配。实验结果证明基于加权IMF对时间序列相似匹配具有很好的效果,能够提高时间序列相似匹配的精度。

参考文献:

- 杨叔子, 吴雅. 时间序列分析的工程应用[M]. 武汉: 华中理工大学出版社, 1991.
- AGRAWAL R, FALOUTSOS C, SWAMI A A. Efficient similarity search in sequence databases[C]//Proc of the 4th International Conference on Foundations of Data Organization and Algorithms. 1993: 13-15.
- MRCHEN F, ULTSCH A. Optimizing time series discretization for knowledge discovery[C]//Proc of the 11th ACM SIGKDD International Conference on knowledge Discovery in Data Mining. New York: ACM Press, 2005: 660-665.
- HUANG N E, SHEN Zheng, LONG S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[C]//Proc of Royal Society of London, Series A. 1998: 903-995.
- MOON Yang-sae, WHANG Kyu-young, LOH Woong-kee. Efficient time-series subsequence matching using duality in constructing windows[J]. Information Systems, 2001, 26(4): 279-289.
- 梁建海, 张建业, 杨峰, 等. 基于斜率偏离的时间序列相似性搜索方法研究[J]. 计算机应用研究, 2010, 27(1): 54-55.
- 吴学雁, 黄道平. 基于事件的时间序列相似性度量方法[J]. 计算机应用, 2010, 30(7): 1944-1946.
- 吴学雁, 黄道平, 莫赞. 基于极值点特征的时间序列相似性查询方法[J]. 计算机应用研究, 2010, 27(6): 2068-2072.
- 张建业, 潘泉, 张鹏. 基于斜率表示的时间序列相似性度量方法[J]. 模式识别与人工智能, 2007, 20(2): 271-274.
- 李海林, 郭崇慧. 基于多维形态特征表示的时间序列相似性度量[J]. 系统工程理论与实践, 2011, 31(4): 993-1004.