

基于 EMD 的时间序列相似性度量算法

贾瑞玉,王 瑞

(安徽大学 计算机科学与技术学院,安徽 合肥 230601)

摘 要:时间序列本身具有高维、高噪声的特点。在进行相似性度量之前,需要对序列进行特征表示。针对时间序列相似性度量工作中,使用分段线性表示方法对序列进行特征表示,分段拟合效果依赖于划分粒度,若分段数和分段点选取不当,可能导致拟合效果不佳、难以准确反映序列整体形态趋势的问题,提出一种新的基于趋势的相似性度量方法。该方法将经验模态分解方法与分段线性表示方法相结合,首先用经验模态分解方法过滤细节信息,提取序列的主要形态趋势,得到趋势拟合序列。在此基础上,再用分段线性表示方法对趋势拟合序列进行分段表示,减少拟合结果对划分粒度的依赖性。最后给出序列的分段向量距离计算方法,对趋势分段序列计算加权向量距离,得到不同序列之间的相似性。仿真实验表明,该算法稳定有效、对噪声不敏感。

关键词:时间序列;相似性;趋势;向量距离

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2017)11-0071-04

doi:10.3969/j.issn.1673-629X.2017.11.015

A Similarity Measure Algorithm for Time Series Based on EMD

JIA Rui-yu, WANG Rui

(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract: The time series itself has characteristics of high dimension and high noise. It is necessary to represent the sequence before the similarity measure. When using piecewise linear representation method for feature representation, piecewise fitting results depend on the partition granularity. If the segmentation number and segmentation points are not proper selection, which may lead to poor fitting and could not accurately reflect the overall trend of the sequence form. Therefore, aiming at the problem, a new method of similarity measurement based on the trend is proposed which combines the empirical mode decomposition with piecewise linear representation. Firstly, filtering details by empirical mode decomposition, extracting main morphological trend of the sequence, the trend fitting sequence is gained. On this basis, use of piecewise linear representation for fitting the trend sequence, the dependence of the fitting result on the partition granularity is reduced. Finally, the calculation method of piecewise vector distance is given. The similarity between different sequences can be obtained by calculating the weighted vector distance of the trend segment sequence. Simulation results show that the proposed algorithm is stable and effective, and not sensitive to noise.

Key words: time series; similarity; trend; vector distance

0 引言

时间序列是一组与时间相关的高维数据,在金融、商业、医学及社会科学等领域中广泛存在,如股市行情、Web访问量、脑电图分析和气象变化数据等^[1-2]。时间序列相似性度量是时间序列的聚类、预测及相似性搜索的基础,所以时间序列的相似性研究具有重要的现实意义和广泛的应用前景^[3]。由于时间序列具有数据量巨大、噪声含量多、短期起伏、非稳态等特点,对

这类数据进行挖掘分析的难度较大,因此时间序列的近似表示与相似性度量成为时间序列数据挖掘的研究热点^[4-5]。

基于分段思想的序列表示方法能较好地保留时间序列的全局特征,降低数据维度,并且计算效率高,因此被广泛应用。研究人员也提出了很多时间序列的相似性度量方法。

欧氏距离具有直观、高效的优点,但对数据的平移

收稿日期:2016-11-03

修回日期:2017-03-15

网络出版时间:2017-08-01

基金项目:国家自然科学基金资助项目(61202227)

作者简介:贾瑞玉(1965-),女,副教授,硕士研究生导师,研究方向为数据挖掘、智能软件;王 瑞(1991-),女,硕士研究生,研究方向为数据挖掘、智能数据处理。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20170801.1550.026.html>

非常敏感,且计算精度不高^[6]。吴虎胜等^[7]提出了一种基于二维奇异值分解的相似匹配方法,其本质是用二维奇异值分解对序列进行特征描述,然后计算欧氏距离。Berndt 和 Chen 等^[8-9]将动态弯曲距离用于时间序列的相似性度量,克服了欧氏距离无法反映数据整体发展趋势的缺点,并且可以处理不同长度的时间序列,但是存在复杂度较高的问题。刘彤彤等^[10]提出了基于窗口斜率表示的相似性算法,以每个窗口内最大最小幅值差与窗口宽度的比值作为序列的特征信息,进行相似性分析。李海林等^[11]提出的基于多维形态表示的时间序列相似性度量方法,利用正交多项式的基向量形成的特征空间对序列进行特征表示,进而计算相似度。但特征空间的选择对相似性度量结果影响较大。董晓莉等^[12]提出的七元模式形态距离虽然保留了时间序列的分段趋势信息,但本质上是基于对序列分段模式的有限划分,因此,任意两个序列对应分段间的距离值都是离散的,相似匹配的精确程度依赖于模式划分粒度。

针对现有方法的一些不足,文中提出一种基于经验模态分解的相似性度量算法。用经验模态分解方法提取时间序列的趋势信息,并在此基础上进行时间序列的分段线性表示,然后用分段向量距离方法计算趋势分段序列的相似性。基于趋势的分段能有效降低噪声影响,加权向量距离从数据趋势上区分差异,比基于点距离的方法更稳定,同时修正了用户间可能存在的度量标准不统一的问题。

1 相关工作

因为时间序列具有数据量巨大、增长速度快、含有大量噪声等特点,因此在分析时间序列数据之前需要进行维度约简。由 Keogh 等^[13]提出的 PLR 分段线性表示方法具有良好的数据压缩作用。PLR 方法的基本思想是选取序列中的局部极值点或拐点,若该点与端点的比值大于参数 ε , 则该点被认为是重要点。将所有重要点用直线连接,即可得到基于重要点的分段时间序列。参数 ε 的大小决定了分段表示的划分粒度。

$S = \langle u_1, u_2, \dots, u_i, \dots, u_n \rangle$ 是原始时间序列,其中 $i = 1, 2, \dots, n$, n 是序列长度, $u_i = (t_i, x_i)$ 是原始时间序列中第 i 个点,表示 t_i 时刻的数据取值为 x_i 。使用 PLR 方法的基本思想是选取序列中最重要的某些点,构成一系列线段,因此把 PLR 分段序列表示为 $S'_1 = \langle u'_1, \dots, u'_j, u'_{j+1}, \dots, u'_m \rangle$ 。其中 u'_j, u'_{j+1} 为第 j 段的起点和终点, m 为重要点的数量。

已知分段序列 $S'_1 = \langle u'_1, \dots, u'_j, u'_{j+1}, \dots, u'_m \rangle$, 定义 S'_1 的向量序列为 $P_s = \langle \vec{p}_1, \dots, \vec{p}_j, \dots, \vec{p}_{m-1} \rangle$ 。其

中, $\vec{p}_j = u'_{j+1} - u'_j$ 是分段序列中分段 j 对应的二维向量,表示分段 j 的方向,反映出当前分段的发展趋势。

2 基于 EMD 的相似度量方法

2.1 EMD 方法提取序列趋势

经验模态分解方法 (Empirical Mode Decomposition, EMD) 是 Hilbert-Huang 等提出的适用于非平稳时间信号分析的方法。它的核心思想是将被分析的数据分解成多个具有不同特征的数据序列组合,每个数据序列具有一个本征模函数 (Intrinsic Mode Function, IMF) 信号。Huang 等定义 IMF 必须满足以下两个条件^[14]:

(1) 序列数据的极值点个数和过零点个数相差不超过 1;

(2) 由极大值点构成的上包络线和由极小值点构成的下包络线关于时间轴对称。

满足以上条件的信号即为 IMF 信号。如果信号不满足上述条件,则需要分解操作,进行平稳化处理,将原始信号分解成多个满足 IMF 的子信号。

若原始序列信号用 $X(t)$ 表示,则分解过程主要分为以下三个步骤:

步骤 1: 找出原始序列中的局部极大值和局部极小值,通过三次样条插值分别得到由极大值构成的上包络线 $U(t)$ 和由极小值构成的下包络线 $L(t)$,将序列中的所有信号点都包含在两条包络线之间,上包络线和下包络线的平均值为 $m_1(t)$ 。原始序列和上下包络线的平均值包络线差值为 h_1 ,如果序列 h_1 不满足 IMF 的条件或者给定阈值,执行步骤 2,否则执行步骤 3。

步骤 2: 将 h_1 作为原始信号重复执行步骤 1,直至执行结果满足 IMF 条件或给定阈值。

$$h_{11}(t) = h_1(t) - m_{11}(t) \quad (1)$$

设重复执行 i 次后,结果满足 IMF 或阈值内的误差。

$$h_{1i}(t) = h_{1(i-1)}(t) - m_{1i}(t) \quad (2)$$

步骤 3: 如果序列 $h_{1i}(t)$ 满足 IMF 条件,则将其表示为:

$$c_1(t) = h_{1i}(t) \quad (3)$$

$$r_1 = X(t) - c_1(t) \quad (4)$$

其中, r_1 为残留分量。将 r_1 作为原始序列重复步骤 1 的筛选过程,直至满足下列条件之一:

- (1) r_n 或者 c_n 小于给定阈值;
 - (2) r_n 成为单调函数,无法再从分解中得到 IMF。
- 最后得到:

$$X(t) = \sum_{i=1}^n C_i(t) + r_n(t) \quad (5)$$

其中, $X(t)$ 表示原始时间序列信号; $C_i(t)$ 表示满足 IMF 的子信号; $r_n(t)$ 为趋势序列。

因为 EMD 分解过程是依据原始数据信息进行的, 因此分解的信号隐含数据的真实信息, $r_n(t)$ 体现了序列的真实趋势。

图 1 为一个经验模态分解过程。其中, Signal 是原始序列信号; imf1-imf5 是分解过程的细节; res 是分解后得到的趋势序列。将 imf 信号累加可以基本拟合原始序列。

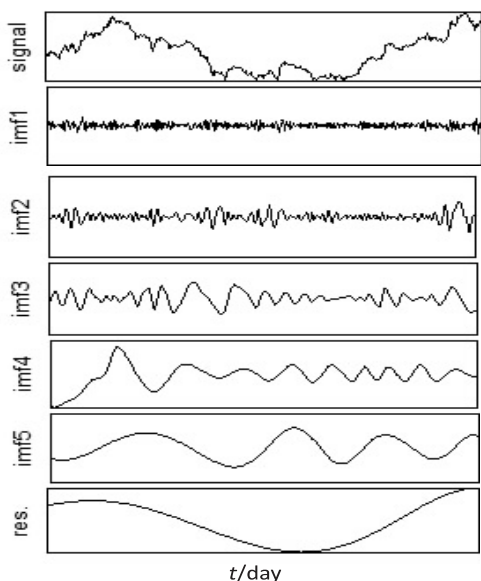


图 1 经验模态分解过程

图 2 为原始序列与分解后的趋势拟合序列对比图。

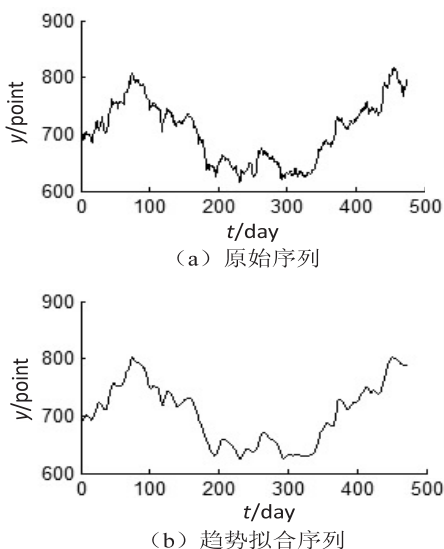


图 2 原始序列与趋势拟合序列对比

直接对原始序列进行分段表示的计算量较大, 而且容易受噪声影响。拟合序列能在保留基本趋势的基础上过滤噪声, 降低数据维度。对拟合序列使用 PLR 分段表示方法, 提取到的分段序列趋势特征更准确, 能够提高相似性度量的准确性。

2.2 基于 EMD 的向量距离

在计算两个时间序列 S_1 和 S_2 之间的距离之前, 先利用 EMD 方法提取趋势序列, 然后进行分段表示, 得到分段序列 $S'_1 = \langle u'_1, \dots, u'_i, u'_{i+1}, \dots, u'_m \rangle$ 和 $S'_2 = \langle v'_1, \dots, v'_i, v'_{i+1}, \dots, v'_m \rangle$ 。转换为向量序列后分别对应于两个等长向量序列 $P = \langle \vec{p}_1, \dots, \vec{p}_i, \dots, \vec{p}_{m-1} \rangle$ 和 $Q = \langle \vec{q}_1, \dots, \vec{q}_i, \dots, \vec{q}_{m-1} \rangle$ 。

每一对分段向量之间的夹角为 $\theta_i = \langle \vec{p}_i, \vec{q}_i \rangle, i = 1, 2, \dots, m-1$ 。为了使相似度和距离成正比, 定义分段向量距离为:

$$s_i = 1 - \cos\theta_i = 1 - \frac{\vec{p}_i \cdot \vec{q}_i}{|\vec{p}_i| |\vec{q}_i|} \quad (6)$$

分段向量距离区间为 $[0, 2]$, 不同分段向量之间夹角越大, 向量距离越大, 相似度越低; 夹角越小, 向量距离越小, 相似度越高。当向量距离为 0 时, 表示两个向量方向完全一致, 即两个分段的趋势完全相同; 向量距离为 2 时, 表示两个分段趋势完全相反。通过分段相似度可以得到两个序列的整体相似度:

$$\text{Sim}(P, Q) = \sum_{i=1}^n (t_{i+1} - t_i) s_i / t_n \quad (7)$$

为了提高计算结果的准确性, 在相似度公式中引入权重 $(t_{i+1} - t_i) / t_n$, 表示分段 i 的相似性 s_i 在整体相似性中所占的权重。其中, t_n 表示整个序列的长度, $t_{i+1} - t_i$ 表示分段 i 的长度。 $(t_{i+1} - t_i) / t_n$ 越大, 对整体相似度的影响越大, 加权后得到的结果能够更精确地反映相似度。

2.3 基于 EMD 的相似性度量算法

输入: 原始时间序列 $S_1 = \langle u_1, u_2, \dots, u_i, \dots, u_n \rangle, S_2 = \langle v_1, v_2, \dots, v_i, \dots, v_n \rangle$ 。

Step1: 使用 EMD 方法对原始序列进行经验模态分解;

Step2: 拟合原始序列的趋势序列;

Step3: 对趋势拟合序列进行分段, 得到分段序列;

Step4: 对不同序列的分段序列进行等长处理;

Step5: 计算分段序列的向量序列;

Step6: 计算不同向量序列之间的相似度。

输出: S_1 和 S_2 的相似度 Sim。

使用 EMD 提取序列趋势之后进行分段表示, 减少了噪声的影响, 能够更准确地反映序列的形态变化特征, 使分段表示提取的趋势特征更准确。向量距离根据不同分段的发展趋势区分差异, 比基于点距离的方法精确, 比基于形态距离的方法灵活。

3 实验

实验选取 Intel Core(TM) i5-3210M CPU @ 2.50

GHz,内存为 4 GB 的电脑,操作系统为 Microsoft Windows7,开发工具为 MATLAB。采用三种股票在 2005 年至 2010 年之间的每日收盘价格时间序列进行实验。第一组数据 S_1 是标普 500 指数数据(S&P500),第二组数据 S_2 是沪深 300 指数数据,第三组数据 S_3 是上证指数数据,每条时间序列包含 1 500 个数据点。图 3 是三种时间序列原始数据的走势图。为防止度量标准不统一对结果造成影响,在预处理过程中会将数据统一映射到[0,1]区间。

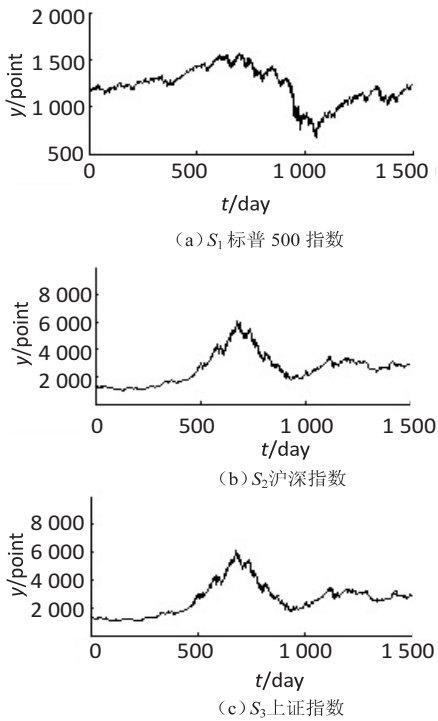


图 3 三种股票数据的走势图

对比方法采用欧氏距离方法和形态距离方法。欧氏距离方法和形态距离方法直接使用 PLR 分段方法,将预处理后的序列分别压缩至 50、100 和 150 段,然后进行距离计算。文中方法先利用 EMD 进行数据分解,拟合原始序列,再使用 PLR 方法对拟合序列进行压缩分段表示,最后使用加权向量距离计算相似度。实验结果如表 1~3 所示。

表 1 文中方法的实验结果

分段数	S_1 与 S_2	S_1 与 S_3	S_2 与 S_3
50	0.064 8	0.065 3	0.027 5
100	0.039 1	0.040 8	0.019 6
150	0.063 5	0.057 6	0.023 4

表 2 形态距离法的实验结果

分段数	S_1 与 S_2	S_1 与 S_3	S_2 与 S_3
50	0.075 1	0.087 5	0.031 2
100	0.072 6	0.065 9	0.041 3
150	0.084 7	0.079 2	0.035 1

表 3 欧氏距离法的实验结果

分段数	S_1 与 S_2	S_1 与 S_3	S_2 与 S_3
50	2.351 3	2.819 6	1.512 6
100	4.384 2	4.032 7	2.191 7
150	5.442 7	3.298 4	3.582 3

实验结果表明,加权向量距离和形态距离方法结果一致,无论将数据压缩到 50 段、100 段还是 150 段,结果均是 S_2 与 S_3 的距离最小,没有受压缩程度的影响,欧氏距离方法在压缩到 150 段时认为 S_1 与 S_3 最相似。对比图 3 中的数据形态走势图可以发现,加权向量距离和形态距离结果正确,但是欧氏距离在 150 段时没有正确区分序列的趋势差异,计算结果出现误差。虽然本次实验中形态距离和向量距离结果均正确,但是形态距离方法的划分粒度不同,度量结果也会有差异,因此不够稳定。基于 EMD 的向量距离方法虽然多了提取序列趋势的步骤,但是减少了分段表示的计算量,而且计算结果稳定有效,适用于序列的相似分析和相关研究。

4 结束语

针对现有方法的一些不足,提出了基于 EMD 的向量距离相似性度量方法。实验结果表明,该方法能够在降低噪声影响的同时提取序列的趋势信息,向量距离能够有效度量时间序列的形态趋势相似度,而且不需要划分形态模式,避免了形态距离方法划分标准不统一、度量结果不稳定的问题,同时克服了欧氏距离对数据平移敏感的缺陷。下一步工作主要是考虑如何减少计算时间,并将该度量方法应用到时间序列的聚类分析中。

参考文献:

[1] 张海涛,李志华,孙雅,等. 时间序列的层次分段及相似性度量[J]. 计算机工程与应用,2015,51(10):147-151.

[2] 朱扬勇,戴东波,熊赞. 序列数据相似性查询技术研究综述[J]. 计算机研究与发展,2010,47(2):264-276.

[3] 涂辉,刘丽,张正金. 改进 DTW 算法的心电信号相似性度量[J]. 计算机工程与应用,2015,51(16):215-218.

[4] 刘永志,皮德常,陈传明. 基于关键点不同长度时间序列相似性度量[J]. 计算机工程与应用,2014,50(20):1-4.

[5] 尚福华,马楠,杜睿山. 基于形态特征的测井曲线相似性搜索研究[J]. 计算机应用研究,2013,30(4):1076-1078.

[6] 张勇,王元珍,曹忠升. 基于形态拟合的时间序列距离计算[J]. 华中科技大学学报:自然科学版,2012,40(8):72-76.

[7] 吴虎胜,张凤鸣,钟斌. 基于二维奇异值分解的多元时间序列相似匹配方法[J]. 电子与信息学报,2014,36(4):847-854.

(下转第 78 页)

和稀疏编码系数同时都具有区分性,故文中算法的字典能很好地重建出模糊车牌中文字符,使得中文车牌字符不易于与其他中文结构相近的省份车牌字符所混淆。

3 结束语

文中提出了一种模糊车牌识别的新方法。该方法采用基于费希尔判决准则的字典学习模型和 Softmax 回归来分别表示和识别模糊车牌中文字符。相比一般的字典学习算法,基于费希尔判决准则的字典学习对于模糊车牌中文字符有着更强的重建能力。此外,整合三个字典模型进行识别的方法比只凭一个识别器识别的方法具有更高的识别率和更好的稳定性。因此,可以广泛应用于车牌系统识别中。

参考文献:

- [1] Lan C, Li F, Jin Y, et al. Research on the license plate recognition based on image processing [C]//Fifth international conference on instrumentation and measurement, computer, communication and control. Qinhuangdao: IEEE, 2015: 731–734.
- [2] 邹明明, 卢迪. 基于改进模板匹配的车牌字符识别算法实现[J]. 国外电子测量技术, 2010, 29(1): 59–61.
- [3] 吕润华, 苏婷婷, 马晓伟. BP 神经网络联合模板匹配的车牌识别系统[J]. 清华大学学报: 自然科学版, 2013(9): 1221–1226.
- [4] 吴聪, 殷浩, 黄中勇, 等. 基于人工神经网络的车牌识别[J]. 计算机技术与发展, 2016, 26(12): 160–163.
- [5] Vishwanath N, Somasundaram S, Nishad A, et al. Indian license plate character recognition using Kohonen neural network [C]//International conference on computational intelligence & computing research. [s. l.]: IEEE, 2012: 1–4.
- [6] Liu P, Li G, Tu D. Low-quality license plate character recog-

nition based on CNN [C]//2015 8th international symposium on computational intelligence and design. Hangzhou: IEEE, 2015: 53–58.

- [7] Zhong Z, Jin L, Feng Z. Multi-font printed Chinese character recognition using multi-pooling convolutional neural network [C]//13th international conference on document analysis and recognition. Tunis: IEEE, 2015: 96–100.
- [8] Bautista R M J S, Navata V J L, Ng A H, et al. Recognition of handwritten alphanumeric characters using projection histogram and support vector machine [C]//International conference on humanoid, nanotechnology, information technology, communication and control, environment and management. [s. l.]: [s. n.], 2015: 1–6.
- [9] Angeline L, Wei Y K, Wei L K, et al. Research of license plate character features extraction and recognition [C]//2nd international conference on computer science and network technology. [s. l.]: [s. n.], 2012: 2154–2157.
- [10] Ghahnavieh A E, Amirkhani-Shahraki A, Raie A A. Enhancing the license plates character recognition methods by means of SVM [C]//22nd Iranian conference on electrical engineering. [s. l.]: [s. n.], 2014: 220–225.
- [11] 周鹏. 基于支持向量机的车牌字符识别方法[J]. 数字技术与应用, 2016(9): 91.
- [12] 陈思宝, 赵令, 罗斌. 基于核 Fisher 判别字典学习的稀疏表示分类[J]. 光电子·激光, 2014, 25(10): 2000–2008.
- [13] 练秋生, 石保顺, 陈书贞. 字典学习模型、算法及其应用研究进展[J]. 自动化学报, 2015, 41(2): 240–260.
- [14] 朱杰, 杨万扣, 唐振民. 基于字典学习的核稀疏表示人脸识别方法[J]. 模式识别与人工智能, 2012, 25(5): 859–864.
- [15] 刘冰冰. 基于 PCA 车牌汉字识别算法的研究与实现[D]. 长春: 长春理工大学, 2011.
- [16] 闫雪梅, 王晓华, 夏兴高. 基于 PCA 和 BP 神经网络算法的车牌字符识别[J]. 激光与红外, 2007, 37(5): 481–484.

(上接第 74 页)

- [8] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series [C]//Working notes of the knowledge discovery in databases workshop. [s. l.]: [s. n.], 1994: 359–370.
- [9] Chen Y, Hu B, Keogh E, et al. DTW-D: time series semi-supervised learning from a single example [C]//ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]: ACM, 2013: 383–391.
- [10] 刘彤彤, 戴敏, 李忠义. 基于窗口斜率表示法的心电波形相似性分析[J]. 计算机应用, 2012, 32(10): 2969–2972.
- [11] 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, 2013, 33(4): 1024–

1034.

- [12] 董晓莉, 顾成奎, 王正欧. 基于形态的时间序列相似性度量研究[J]. 电子与信息学报, 2007, 29(5): 1228–1231.
- [13] Keogh E J, Pazzani M J. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback [C]//International conference on knowledge discovery & data mining. [s. l.]: [s. n.], 1998: 27–31.
- [14] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis [J]. Proceedings of the Royal Society A Mathematical Physical & Engineering Sciences, 1998, 454(1971): 903–995.