基于异常序列剔除的多变量时间序列结构化预测

毛文涛 [1,2,3] 蒋梦雪 [1] 李 源 [1,2] 张仕光 [1,2]

摘 要 针对传统多变量时间序列预测方法未考虑变量间依赖关系从而影响预测效果的问题, 提出了一种基于异常序列剔除的多变量时间序列预测算法. 该算法旨在利用多维支持向量回归机 (Multi-dimensional support vector regression, M-SVR) 内在的结构化输出特性, 对选取到具有相似性的多个变量序列进行联合预测. 首先, 对已知序列进行基于模糊熵的层次聚类, 实现对相似序列的初步划分; 其次, 求出类中所有序列的主曲线, 根据序列到主曲线的距离计算各个序列的异常因子, 从而进一步剔除聚类结果中的异常序列; 最后, 将选取到的相似变量序列作为输入, 利用 M-SVR 进行预测. 通过理论分析, 证明本文算法在理论上存在信息损失上界与可靠度下界, 从而说明本文算法的合理性与可行性. 采用混沌时间序列数据与多个实际数据集进行对比实验, 结果表明, 与现有多个代表性方法相比, 本文算法可有效挖掘多变量时间序列的内在结构信息, 预测精度更高, 数值稳定性更好.

关键词 时间序列聚类, 主曲线, 异常序列, 多维支持向量回归机

引用格式 毛文涛, 蒋梦雪, 李源, 张仕光. 基于异常序列剔除的多变量时间序列结构化预测. 自动化学报, 2018, **44**(4): 619−634

DOI 10.16383/j.aas.2017.c160707

Structural Prediction of Multivariate Time Series Through Outlier Elimination

MAO Wen-Tao[1, 2, 3] JIANG Meng-Xue[1] LI Yuan[1, 2] ZHANG Shi-Guang[1, 2]

Abstract To solve the problem that the traditional multivariate time series prediction generally ignores the dependency among all variables, a new multivariate time series structural prediction method through outlier elimination is proposed. This algorithm predicts on the selected multivariate time series by using the structural output characteristic. Firstly, to recognize the relatedness among the sequences, the variable sequences are initially divided by hierarchical clustering according to fuzzy entropy. Secondly, to further evaluate the similarity of the sequences in the obtained cluster, the principal curve is introduced to calculate the abnormality degree of each sequence, and then the outlier sequence can be eliminated in terms of the value of abnormality degree. As a result, similar sequences can be distinguished. Finally, for the similar series, multi-dimensional support vector regression (M-SVR) is used to construct the prediction model, and then the structural prediction for multivariate time series is conducted. Moreover, a theoretical proof is provided to show the proposed method has an upper bound of the loss of information and a lower bound of reliability and that the proposed method is reasonable and feasible from the perspective of information entropy. Experiments are conducted on three chaotic time series datasets and five real-life datasets. The results show that the proposed method can effectively recognize the inner group structure among multivariable sequences, so as to obtain a better forecasting accuracy and numerical stability than those widely used methods in terms of two different error measurements.

Key words Time series clustering, principal curve, outlier sequence, multi-dimensional support vector regression (M-SVR)

Citation Mao Wen-Tao, Jiang Meng-Xue, Li Yuan, Zhang Shi-Guang. Structural prediction of multivariate time series through outlier elimination. *Acta Automatica Sinica*, 2018, **44**(4): 619−634

在实际应用中, 时间序列数据的变化往往受多种因素影响. 传统的时间序列预测算法主要针对影响事物变化的某一种因素进行单个时间变量的预测.

收稿日期 2016-10-10 录用日期 2017-02-07
Manuscript received October 10, 2016; accepted February 7, 2017
国家自然科学基金 (U1204609), 中国博士后科学基金 (2016T90944), 河南省高校科技创新人才资助计划 (15HASTIT022), 河南省高校青年骨干教师资助计划 (2014GGJS-046), 河南师范大学优秀青年科学基金 (14YQ007), 河南省高等学校重点科研项目 (16A520015) 资助
Supported by National Natural Science Foundation of China (U1204609), China Postdoctoral Science Foundation (2016T90944), the Funding Scheme of University Science and Technology Innovation of Henan Province (15HASTIT022), the Funding Scheme of University Young Core Instructor of Henan Province (2014GGJS-046), the Foundation of Henan Normal University for Excellent Young Teachers (14YQ007), and the Key Research Project of High School of Henan Province (16A520015)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling

1. 河南师范大学计算机与信息工程学院 新乡 453007 2. 河南省高校计算智能与数据挖掘工程技术研究中心 新乡 453007 3. 西北工业大学力学与土木建筑学院 西安 710129
1. College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007 2. Computational Intelligence and Data Mining Engineering Technology Research Center of Colleges and Universities of Henan Province, Xinxiang 453007 3. School of Mechanics and Civil & Architecture, Northwestern Polytechnical University, Xi'an 710129

而若对事物变化趋势进行更加综合、准确的评估, 则需对多个相关因素进行同时预测, 即多变量时间序列预测. 相比于单变量时间序列预测, 多变量时间序列预测可同时预测多个变量的走势, 同时可利用变量之间的相关信息提高动态预测的精度与稳定性, 已得到越来越多学者的关注. 但是, 传统预测方法直接应用于多变量时间序列预测中, 容易受变量之间冗余作用、误差累积和缺乏关联信息等特点的影响, 无法取得令人满意的预测效果. 因此, 选择一个合适的建模方法对多变量进行更为准确地预测有着重要的理论价值和现实意义.

目前, 多变量时间序列预测研究已取得一定的进展. 传统的预测方法将多变量分解为多个单变量, 利用支持向量回归机 (Support vector machine, SVM) 等方法[1] 对每个变量单独进行回归建模. 这种方法简单明了, 但是重复建模增加了计算量, 同时未能有效利用变量之间的结构信息. 一种典型的改进方法是利用时间序列本身的系统特性信息和数据特点进行预测. 其中, 张勇等[2] 利用最大 Lyapunov 指数理论, 选取多个邻近参考点, 实现了对多个混沌序列的同时预测; 针对股指期货价格预测问题, Sun 等[3] 利用模糊 C 均值对数据进行预处理, 并结合粗糙集算法建立模糊逻辑关系组, 实现对多个变量走势的预测; 韩敏等[4] 利用主成分分析方法实现对输入变量降维, 再将动态储备迟用作核函数充分映射多元混沌时间序列的动力学特性实现了避免过拟合, 提高预测精度. 另一种做法是围绕着多变量输出结构的算法改进. 例如, Wang 等[5] 提出基于极限学习机的在线多变量时间序列预测方法, 对多变量序列进行相空间重构后, 建立极限学习机预测模型, 实现在线预测; Han 等[6] 提出基于 SCKF-γESN 模型的在线多变量时间序列预测方法, 利用平方根容积卡尔曼滤波算法更新 γ 回声网络参数, 以实现对多变量序列的在线预测, 同时在滤波算法中加入异常值的检测, 使得预测结果更加稳定; Chen 等[7] 利用 K 近邻和互信息获取多变量时序数据的重要性表示, 并据此重构样本, 利用改进的加权 LS-SVM 进行预测. 但是, 这些方法大多通过多变量数据的简单融合构建预测模型, 缺乏在算法结构上针对多变量序列特点的有效改进.

由上述分析可知, 对多变量时间序列预测来说, 预测效果的好坏直接取决于数据中蕴含的有效信息量. 而实际应用多为小样本预测问题, 若能有效利用变量之间的结构化信息, 则可一定程度弥补样本数不足带来的信息缺失, 有利于提高小样本下预测模型的稳定性和精度. 因此, 提升多变量时间序列预测效果的关键在于: 1) 如何有效挖掘序列之间的依赖关系等结构化信息? 2) 如何构建适用于多变量序列预测的预测模型? 针对这两点, 本文提出一种基于异常序列剔除的多变量时间序列预测方法. 该方法首先给出基于模糊熵的时间序列聚类算法, 实现对相似序列的初步划分, 并引入主曲线, 构建异常时间序列检测算法, 对其中的异常序列进行剔除, 最后采用多维度支持向量回归机 (Multi-dimensional support vector regression, M-SVR)[8] 对最终得到的相似序列进行多输出时间序列预测. M-SVR 是一种具有多输出结构的支持向量机, 不仅对小样本有快速和准确的回归预测效果, 而且, 利用超球损失函数度量多个输出端的的风险损失, 可有效利用输出端之间的结构化信息, 目前已在多步超前时间序列预测[9] 等问题取得成功应用. 但根据作者文献调研, 尚未发现 M-SVR 在多变量时间序列预测中的应用. 鉴于此, 本文采用 M-SVR 作为基础建模算法, 旨在利用 M-SVR 的结构化输出特性, 选择具有相关性的序列同时进行预测, 以达到更好的预测效果. 此外, 本文从理论上给出了异常序列剔除的信息损失上界和模型可靠度下界, 从而证明所提算法的合理性. 最后采用混沌时间序列数据与五个实际数据集数据进行仿真实验, 实验结果验证了所提算法的有效性.

1 相关理论

1.1 层次聚类方法

层次聚类方法是一种常用的聚类方法, 分为凝聚层次聚类方法和分裂层次聚类算法. 该类算法的核心思想是递归地采用自底向上或自顶向下策略对数据对象进行合并或分裂. 分裂层次方法是把一个给定的数据对象族迭代地分裂, 进而形成更小的数据对象. 该类算法的优点是能够获取到不同粒度的多层次聚类结构, 但是算法的复杂度至少为 $\mathrm{O}(n^2)$. 而凝聚层次方法从每一个数据对象开始, 迭代地进行合并, 形成更大的数据对象簇. 目前使用的层次聚类方法多为凝聚层次聚类, 中间的区别一般是定义的类间距离不同. 此外, 该类算法需要事先给定一个族合并或分裂的终止准则, 并且某个族一旦完成合并或分裂步骤就无法撤消[10]. 层次聚类是一种嵌套聚类的方法, 通用性强, 适用于小数据集[11].

1.2 时间序列的相似性度量

时间序列的相似性度量是衡量两个时间序列相似的标准, 其有效性直接关系到后续工作的性能. 目前对时间序列相似性的度量主要有两种方法.

1) 欧氏距离通过序列中对应值计算得到对应的相似性, 其计算公式为

$$D(X,L)=2\sqrt{\sum_{i=1}^{n}(x_i-l_i)^2}$$

其中, i 为序列中的第 i 个值[12]. 时间序列的欧氏距离需要两个序列等长, 且要求序列的值一一对应.

2) 动态时间弯曲 (Dynamic time warping distance, DTW) 距离[13] 是一种通过弯曲时间轴来更好地对时间序列形态进行匹配映射的相似性度量方法. DTW 距离在两个时间序列之间寻找最小的映射路径, 允许一个序列中的点对应于另一个序列中的多个相邻的点. DTW 不需要利用领域知识, 只是假设相近的序列存在低消耗的平移匹配, 并在时间序列分类上取得了广泛应用[14]. DTW 距离是一种基于距离测量指标的算法, 其基本假设前提为数据符合正态分布. 设序列 s 的长度是 M, 序列 c 的长度是 L, 两个时间序列的 DTW 距离计算过程如下: 首先计算矩阵中的 (i,j) 两点之间的欧氏距离: $d(s_i,c_j)=\|s_i-c_j\|^2$, 得到一个 $M\times L$ 的距离矩阵, 然后将两个序列的规整路径定义为 $W=(w_1,\cdots,w_k)$, 其中 $w_k=(i,j)_k=d(s_i,c_j)$. DTW 的目的是找到一个从 $(1,1)$ 到 (M,L) 使 $\sum_{k=1}^{K}w_k$ 取到最小值的单调增长路径[13]. 最佳路径是 $\gamma_{i,j}=d(s_i,c_j)+\min\{\gamma_{i-1,j-1}+\gamma_{i-1,j}+\gamma_{i,j-1}\}$.

1.3　主曲线

主曲线是通过数据集 “中间” 的光滑无参数曲线, 基于一定概率分布下曲线的 “自相合” 特性, 能将数据信息保持好, 有效勾勒出原始信息的轮廓[15], 目的是找到一条通过数据分布 “中央” 并能够真实反映数据的形态的曲线, 这意味着此曲线是数据集合的 “骨架”, 数据集合是这个曲线的 “云”, 所以主曲线对数据的信息保持性好. 具体步骤如下[16]:

步骤 1. 令初始曲线 $f^{(j)}(\lambda)$ 为 X 的第一主成分, 设 $j=0$.

步骤 2. 对所有的 $x\in\mathbf{R}^d$, 求投影指标

$$\lambda_{f^{(j)}}(x)=\max\{t:\left\|x-f^{(j)}(\lambda)\right\|=\min_{\tau}\left\|x-f^{(j)}(\tau)\right\|\}$$

步骤 3. 定义在 x 上 f 的投影点为

$$f^{(j+1)}(\lambda)=\mathrm{E}\left[X|\lambda_{f^{(j)}}(X=\lambda)\right]$$

步骤 4. 如果 $1-\Delta(f^{(j+1)})/\Delta(f^{(j)})$ 小于某个阈值, 则停止 (其中 $\Delta(f^{(j)})$ 表示点 x 到曲线 f 的欧氏平方距离), 否则令 $j=j+1$, 转步骤 2.

2　基于异常时间序列检测的多变量时间序列预测

为提高多变量时间序列预测的准确度和稳定性, 本文提出一种基于异常时间序列检测的多变量时间序列预测方法, 步骤如下:

步骤 1. 计算各条时间序列的模糊熵, 通过比较模糊熵值判断得到序列之间的相似度, 选择层次聚类算法进行聚类以得到初步的相似序列;

步骤 2. 对各类别中的序列计算主曲线, 根据各个序列到主曲线的距离求得异常因子, 并据此进一步剔除聚类结果中相似性相对较弱的序列;

步骤 3. 将得到的相似序列作为输入, 利用 M-SVR 进行多输出建模.

整体算法流程如图 1 所示.

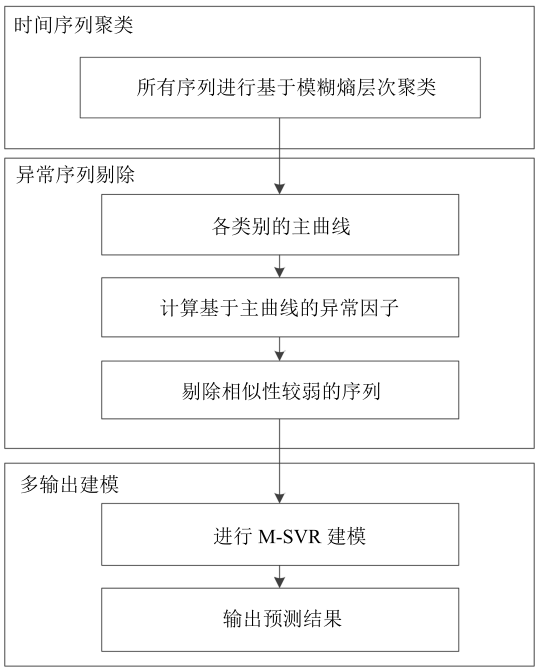


图 1　算法流程图

Fig. 1　The flowchart of the algorithm

2.1　定义

对已知序列定义如下: $\{X_i\}$, $i=1,\cdots,M$, 其中, X_i 为 $\{x_j\}$, $j=1,\cdots,N$. M 表示序列个数, N 代表样本数.

定义 1. 剔除序列过程中序列权重: 构建已知序列集 D 的主曲线, 则 D 中每条序列 x_i 对应的序列权重定义为

$$w_i=1-\frac{f_i}{\sum_{j=1}^{N}f_j}\tag{1}$$

其中, f_i 为序列 x_i 到主曲线的 DTW 距离. 由式 (1) 可知, 序列 x_i 到主曲线的距离越小, 其对应的序列权重越大, 反之 f_i 越大, w_i 越小.

定义 2. 序列的 CP- 距离: 通过距离参数衡量各序列在整个序列挑选过程中的相似度.

$$Sequence(x_i)=f_i\times D_{\mathrm{Fuzz}}(x_i,o_i)\tag{2}$$

其中, f_i 为序列 x_i 到主曲线的 DTW 距离, o_i 为聚类中心, $D_{\mathrm{Fuzz}}(x_i,o_i)$ 为序列到聚类中心的距离. 由

式 (2) 可知, 当序列到聚类中心和主曲线的距离越小, 其对应的 CP- 距离越小.

定义 3. 整个序列挑选过程中序列权重: 针对聚类和主曲线的内容可知, 序列权重跟序列到主曲线的 DTW 距离成反比, 与序列到聚类中心的距离成反比, 定义如下:

$$w_i' = 1 - \frac{f_i \times D_{\text{Fuzz}}(x_i, o_i)}{\sum_{j=1}^{N} f_j} \tag{3}$$

其中, N 和 $\sum_{j=1}^{N} f_j$ 分别为序列数和序列到主曲线距离之和.

2.2 基于模糊熵的时间序列聚类算法

为有效识别序列之间的相似程度, 本文首先对时间序列进行初步的相似性划分. 基本思想是利用时间序列的模糊熵的大小作为衡量序列之间相似性的标准, 对时间序列进行凝聚层次聚类, 建立初始模型. 时间序列凝聚层次聚类将初始序列的每个对象看作一类, 根据相似性度量准则计算类间距离进行合并, 直至满足终止条件. 该方法不需首先确定聚类中心, 可以充分利用类间距离将相对相似的序列聚成一类. 模糊熵采用模糊隶属度函数计算向量之间相似度, 从而使得熵值连续平稳, 对于不同的参数结果稳定, 抗噪性强[17], 因此适合用做层次聚类的类间距. 该算法步骤如下:

步骤 1. 将已知各序列视为一类, 计算每两个序列之间模糊熵的差值作为类间距离.

步骤 2. 根据步骤 1 中得到的类间距离将距离最近的两类合并成一类.

步骤 3. 用 average-linkage 衡量两类 A 与 B 之间的距离:

$$D_{AB} = \frac{1}{|A| \times |B|} \sum_{X_i \in A} \sum_{X_j \in B} D_{\text{Fuzz}}(X_i, X_j)$$

其中, $1 \le i, j \le M$, $i \ne j$, $|A|$ 和 $|B|$ 分别代表类 A 及类 B 的大小.

步骤 4. 重复步骤 2 和步骤 3, 直到满足事先给定的聚类类别数, 即聚类完成.

由上述过程可知, 该时间序列聚类算法采用了模糊熵作为类间距离的衡量方式, 可在整体上有效度量时间序列的相似度, 从而实现对相似序列的初步划分.

2.3 基于主曲线的异常序列的检测

从实验结果可以看出, 在第 2.2 节得到的相似序列中, 往往会因为人为设置类别初始数目等原因, 将一些相似性较弱的序列强制聚到某一类中. 本文将这类序列定义为聚类结果中的异常序列. 异常序列的存在会减弱序列之间的相似度, 从而降低 M-SVR 的结构化预测效果. 因此, 有必要对初步的聚类结果进行进一步的筛选, 剔除相似性相对较弱的异常序列.

根据第 1.3 节分析, 主曲线具有可有效勾勒出原始信息的轮廓和对数据信息保持性好的优点[15]. 因此, 本文在初始阶段聚类的结果上, 通过计算类中序列数据的均值, 找到距离均值距离最近的两条序列, 求得这两条序列的主曲线作为本类的主曲线.

序列集合中, x 序列到主曲线 S 的可达距离定义为

$$RD_k(x, S) = \max\left(\|x - x^{(k)}\|, \|x - S\|\right) \tag{4}$$

其中, $x^{(k)}$ 表示样本中距离序列 x 的 k 近邻样本, $\|x - S\|$ 指从 x 到 S 的 DTW 距离. 基于式 (4), x 的局部可达密度为

$$LRD_k(x) = \left(\frac{1}{k}\sum_{i=1}^{k} RD_k\left(x^{(i)}, x\right)\right)^{-1}$$

x 局部异常因子可定义为 $x^{(i)}$ 的局部可达密度的平均值与 x 局部可达密度的比, 即

$$LOF_k(x) = \frac{\frac{1}{k}\sum_{i=1}^{k} LRD_k\left(x^{(i)}\right)}{LRD_k(x)} \tag{5}$$

$LOF_k(x)$ 的值越大, x 的异常度就越大. 当 $x^{(i)}$ 的周围密度较高, 而 x 的周围密度较低时, 局部异常因子变大, x 被视为异常序列. 相反, 当 $x^{(i)}$ 的周围密度较低, 而 x 的周围密度较高时, 局部异常因子减小, x 被视为正常序列. 因此, 可以根据 $LOF_k(x)$ 的值, 从聚类结果中将异常序列进行剔除. 检测异常值在统计学上已有多种方法, 本文借鉴置信区间的概念, 采用如下方法: 若 $LOF_k(x) > \text{mean}(X) + n \times \text{std}(X)$, 则视为异常序列, 其中 $\text{mean}(X)$ 为序列集合 X 的局部异常因子的均值, $\text{std}(X)$ 为其标准差. 由于此处要剔除的通常为局部异常因子明显较大的序列, 因此在本文实验中, n 直接设为 0.5 即可有效识别各数据集中的异常序列.

2.4 M-SVR 建模

通过以上步骤得到具有依赖关系的相似序列, 根据这些序列构建输入样本集.

$$X(n) = (x(n), x(n+\tau), \cdots, x(n+(m-1)\tau)) \tag{6}$$

其中, $n = 1, 2, \cdots$. 相应序列的下一时刻值作为输出 $Y_i = (x(n+m\tau))$, 其中 m 为嵌入维, τ 为时延, 利用 M-SVR 实现对相似序列的多输出回归建模. 其中 M-SVR 优化目标为[18]

$$Lp = \min_{W,b}\left[\frac{1}{2}\sum_{j=1}^{Q}\left\|w^j\right\|^2 + C\sum_{i=1}^{n}L\left(u_i\right)\right] \tag{7}$$

其中, $L(u) = \begin{cases}0, & u < \varepsilon\\ u^2 - 2u\varepsilon + \varepsilon^2, & u \geq \varepsilon\end{cases}$, $u_i = \|e_i\| = \sqrt{e_i^{\mathrm{T}}e_i}$, $e_i^{\mathrm{T}} = y_i^{\mathrm{T}} - \phi^{\mathrm{T}}(x_i)W - b^{\mathrm{T}}$, $\phi(\cdot)$ 为核映射, $W = [w^1, \cdots, w^Q]$, $b = [b^1, \cdots, b^Q]^{\mathrm{T}}$. 强调输出端之间越相似这个损失项就越小.

利用一阶泰勒展开式可得到其二次近似[18]:

$$Lp'\left(W, b\right) = \frac{1}{2}\sum_{j=1}^{Q}\left\|w^j\right\|^2 + \frac{1}{2}\sum_{i=1}^{n}a_i u_i^2 + CT \tag{8}$$

其中, $a_i = \begin{cases}0, & u_i^k < \varepsilon\\ 2C\left(u_i^k - \varepsilon\right)/u_i^k, & u_i^k \geq \varepsilon\end{cases}$, k 为迭代次数, CT 为独立的高阶项.

通过利用加权迭代最小二乘法 (Iteratively reweighted least squares, IRWLS) 连续迭代过程中线性搜索变量的下降方向, 并收敛到最优值. 由于 $w^j = \sum_i \varphi\left(x_i\right)\beta^j = \phi^{\mathrm{T}}\beta^j$, 以上求解可转换为对参数 β 和 b 的求解, 过程如下[18]:

步骤 1. 设置初值 $k = 0$, $\beta^s = 0$, $b^k = 0$, 计算 u_i^k 和 a^i.

步骤 2. 计算 β^s 和 b^s.

步骤 3. 采用线性搜索, 计算 β^{k+1} 和 b^{k+1}, 并得到 u_i^{k+1} 和 a_i, 判断是否收敛至最优值. 若不收敛, 返回步骤 2.

利用上述过程得到输出端的参数, 可构建得到相应的回归模型.

3　理论分析

本文所提方法旨在利用 M-SVR 的结构特性, 选择具有相似性的多变量时间序列同时进行预测, 关键在于根据序列的异常因子大小进行异常序列的剔除, 得到最具相关性的序列. 为说明该方法的合理性, 本文从信息熵的角度证明在异常序列检测剔除的过程中的信息损失存在上界[19]. 同时, 为证明本算法的有效性, 本文从模型可靠度[20] 入手证明在整个序列挑选的过程中模型的可靠度存在下界.

3.1　异常序列剔除过程信息损失

设剔除序列集合为 $\Psi = \{(x_j, t_j),\ j = 1, 2, \cdots, M - m\}$, 其中 x_j 对应的序列权重为 w_j, 则剔除序列集 Ψ 的总体序列权重之和为

$$\sum_{k=1}^{M-m}w_k = \sum_{k=1}^{M-m}\left(1 - \frac{d_k}{\sum\limits_{j=1}^{M}d_j}\right) \tag{9}$$

易知, $\sum_{j=1}^{M}d_j$ 表示序列集 Φ_d 中所有序列到主曲线的 DTW 距离之和, 对已知序列来说, $\sum_{j=1}^{M}d_j$ 为定值, 故令 $\sum_{j=1}^{M}d_j = \Delta_1$, 则 Ψ 的总体序列损失权重之和为

$$\sum_{k=1}^{M-m}w_k = \sum_{k=1}^{M-m}\left(1 - \frac{d_k}{\Delta_1}\right) = (M-m) - \frac{1}{\Delta_1}\sum_{k=1}^{M-m}d_k \tag{10}$$

定理 1. 令 $H(\Psi)$ 表示异常序列剔除过程中的整体信息损失, 那么有

$$H(\Psi) \leq \left((M-m) - \frac{1}{\Delta_1}\sum_{k=1}^{M-m}d_k\right)\times \log_2\frac{M-m}{(M-m) - \frac{1}{\Delta_1}\sum\limits_{k=1}^{M-m}d_k} \tag{11}$$

其中, 整体信息损失大小 $H(\Psi)$ 的上界仅与剔除序列集 Ψ 中所有序列到主曲线的 DTW 距离之和 $\sum_{k=1}^{M-m}d_k$ 有关.

证明. 根据熵的定义, 有

$$H\left(\Psi\right) = -\sum_{i=1}^{M-m}w_i\log_2 w_i$$

根据最大熵原理, 当每一个 w_i 都取相同的值 $((M-m) - \frac{1}{\Delta_1}\sum_{k=1}^{M-m}d_k)/(M-m)$ 时, $H(\Psi)$ 达到最大值. 则有

$$H\left(\Psi\right) \leq -\sum_{i=1}^{M-m}\frac{(M-m) - \frac{1}{\Delta_1}\sum\limits_{k=1}^{M-m}d_k}{M-m}\times \log_2\frac{(M-m) - \frac{1}{\Delta_1}\sum\limits_{k=1}^{M-m}d_k}{M-m} = \left((M-m) - \frac{1}{\Delta_1}\sum_{k=1}^{M-m}d_k\right)\times \log_2\frac{M-m}{(M-m) - \frac{1}{\Delta_1}\sum\limits_{k=1}^{M-m}d_k} \tag{12}$$

□

由式 (12) 可以看出, $H(\Psi)$ 上界仅和剔除的序列到主曲线的 DTW 距离有关, $\sum_{k=1}^{M-m}d_k$ 越大, 该界越小.

定理 1 从熵的角度给出异常序列剔除过程中的信息损失存在上界, 在理论上证明了根据序列到主曲线的 DTW 距离剔除异常序列样本的可行性与有效性. 考虑极端情况, 若剔除的序列到主曲线的 DTW 距离趋近于无穷大 (即极端不相似), 则对应的信息损失上界趋近于无穷小, 这意味着该序列对整体信息几近可忽略不计, 进一步证明了本文所提异常序列检测方法的合理性.

3.2 模型可靠度

根据上述对算法的描述可知, 假设已知序列 $\{X_i\}$, $i=1,\cdots,M$, 其中, X_i 为 $\{x_j\}$, $j=1,\cdots,N$. M 表示序列个数, N 代表样本数. 挑选之后序列从原来的 $\{X_i\}$, $i=1,\cdots,M$ 到选取得到的 $\{X_j\}$, $i=1,\cdots,n$. 整体摒弃的序列集为 $\{X_t\}$, $t=1,\cdots,M-n$. 已知序列 x_i 的权重为 w_i, 损失序列集合的总权重之和为

$$w'=\sum_{i=1}^{M-n}w_i'=\sum_{i=1}^{M-n}\left(1-\frac{f_i\times D_{\text{Fuzz}}(x_i,o_i)}{\sum\limits_{j=1}^{M-n}f_j}\right)=(M-n)-\sum_{i=1}^{M-n}\left(\frac{f_i\times D_{\text{Fuzz}}(x_i,o_i)}{\sum\limits_{j=1}^{M-n}f_j}\right)\tag{13}$$

则预测值相较于真实值的偏离率

$$p=\sum_{i=1}^{M-n}w_i'=(M-n)-\sum_{i=1}^{M-n}\left(f_i\times\frac{D_{\text{Fuzz}}(x_i,o_i)}{\sum\limits_{j=1}^{M-n}f_j}\right)$$

预测值与真实值的差值

$$E=X_i\times p=X_i\times\sum_{i=1}^{M-n}w_i'=X_i\times(M-n)-X_i\times\sum_{i=1}^{M-n}\left(f_i\times\frac{D_{\text{Fuzz}}(x_i,o_i)}{\sum\limits_{j=1}^{M-n}f_j}\right)$$

已知阈值 θ 的衡量预测效果, 则预测值与真实值的差值 E 超出 θ 时为预测效果不理想的一类 L, 综上可知, L 与 p 呈正相关关系.

定理 2. 由上述描述可知, 此时模型的结果服从二项分布, 在给定置信度 α 的条件下, 模型可靠度的求解公式为

$$\sum_{r=0}^{L}\left(\frac{N}{r}\right)R_L^{N-r}(1-R_L)^r=1-\alpha\tag{14}$$

则模型的可靠度 R_L 存在下限, 且仅与序列 CP- 距离有关.

证明. 由于已知序列个数固定, 聚类之后聚类中心固定, 序列到对应主曲线的投影距离和为定值, 已知

$$\sum_{j=1}^{M-n}d_j\le\sum_{j=1}^{M}d_j=S_1$$

则

$$p=\sum_{i=1}^{M-n}w_i'=(M-n)-\sum_{i=1}^{M-n}\left(\frac{f_i\times D_{\text{Fuzz}}(x_i,o_i)}{\sum\limits_{j=1}^{M-n}f_j}\right)\le(M-n)-\sum_{i=1}^{M-n}\left(\frac{f_i\times D_{\text{Fuzz}}(x_i,o_i)}{\sum\limits_{j=1}^{M}f_j}\right)=(M-n)-\frac{1}{S_1}\sum_{i=1}^{M-n}(f_i\times D_{\text{Fuzz}}(x_i,o_i))$$

即

$$p\le(M-n)-\frac{1}{S_1}\sum_{i=1}^{M-n}(f_i\times D_{\text{Fuzz}}(x_i,o_i)\tag{15}$$

根据可靠度的定义式 (14) 可知, 当 α 确定时, L 与 R_L 呈负相关, 即 p 与 R_L 呈负相关, p 存在上限的同时可靠度 R_L 存在下限, 且结合式 (15) 可知, 可靠度 R_L 仅与序列的 CP- 距离有关, 该值越小, L 越小, 可靠度 R_L 越高, 即若序列与聚类中心、主曲线的距离越近, 则对应的模型可靠度越高, 进一步表明本文所提算法的有效性. □

4 仿真与实验分析

为验证本文所提方法的有效性, 分别引入混沌时间序列数据与五个实际数据集数据进行对比实验. 为方便起见, 本文所提方法简称为 OE-MSVR (Outlier eliminating multi-dimensional SVR).

为进行定量比较, 本文引入均方根误差 (Root mean square error, RMSE) 和平均绝对值误差 (Mean absolute error, MAE) 衡量预测结果. 对应的表达式为

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{pre(i)} - y_{rea(i)}\right)^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(y_{pre(i)} - y_{rea(i)}\right)\right| \tag{16}$$

4.1　混沌时间序列预测

一般认为, 混沌时间序列广泛存在于气象、天文和水文等领域, 具有微观无序、宏观有序的特点. 混沌预测方法就是在相空间中找到一个非线性模型去逼近系统动态特性. 现有的预测方法多为只针对单独序列进行预测. 若能利用混沌序列之间的结构特性实现对多条序列的同时预测, 则可以提高预测效率, 增强预测效果, 更好地把握数据的下一步走势.

对比算法采用超限学习机 (Extreme learning machine, ELM) 和 LibSVM. ELM 是一种单层前馈式神经网络算法, 具有良好的多输出回归能力[21]; LibSVM 是目前广泛使用的 SVM 工具包, 需对每个输出单独建模[22]. 实验中, OE-MSVR 和 LibSVM 均使用 RBF 核函数, 核参数设为 5, 正则化参数 C 均为 100, 松弛变量 ε 为 0.01; ELM 隐神经元个数为 400, 激活函数为 hardlim. 实验开始前, 所有的样本均归一化至 $[-1, 1]$.

4.1.1　混沌时间序列数据生成

本文采用三种典型的混沌时间序列 Lorenz[23]、Mackey-Glass[24] 和 Henon[25] 系统进行实验.

利用上述三种混沌序列构造 8 条时间序列. 其中 Lorenz 序列 2 条, 序号为 1, 2, 分别采用参数 $\sigma = 10$, $r = 28$, $b = 8/3$, 初始值为 $y = [5, 5, 15]$, 如图 2 所示. Mackey-Glass 序列 4 条, 序号为 3, 4, 5, 6, 分别采用参数 $a = 0.2$, $b = 0.1$, $TAU = 25$, 如图 3 所示. Henon 序列 2 条, 序号为 7, 8, 分别采用参数 $a = 1.4$, $x_0 = 0.03$, $y_0 = 0.02$, 如图 4 所示.

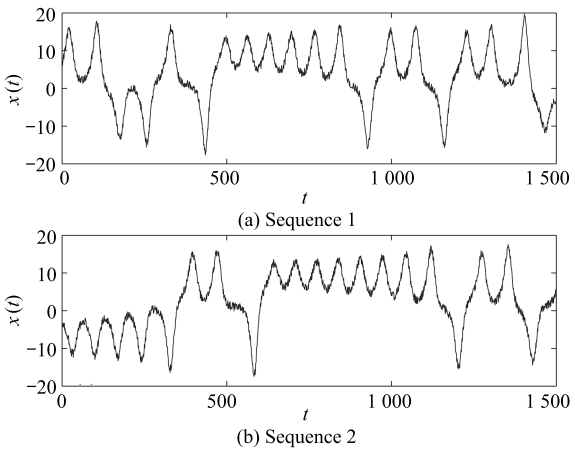


图 2　Lorenz 序列数据

Fig. 2　The Lorenz sequences

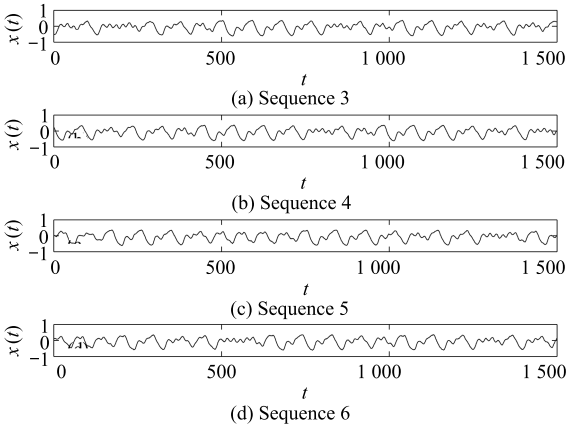


图 3　Mackey-Glass 序列数据

Fig. 3　The Mackey-Glass sequences

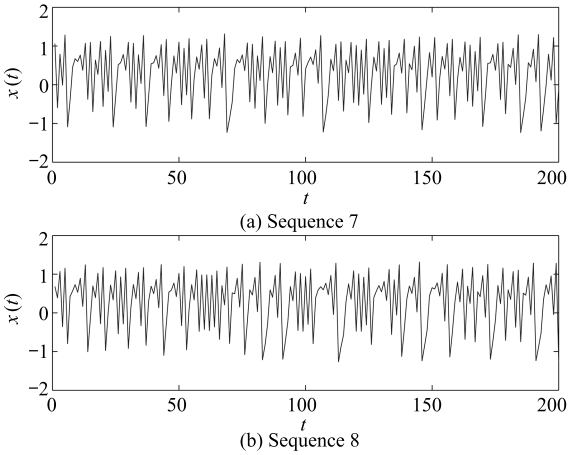


图 4　Henon 序列数据

Fig. 4　The Henon sequences

从几何形态看, Henon 序列与其他两种序列差别较为明显. 每一条时间序列分别产生 1 500 个样本, 其中前 1 000 个样本训练, 后 500 个样本用作测试. 为使得实验场景更接近真实数据, 这里在数据中加入均值为 0、方差为 1 的白噪声.

4.1.2　实验结果与分析

首先, 通过聚类方法得到初步的相似序列集. 考虑到已知序列的数量较少, 为模拟实际应用场景, 将聚类的初始类别数设置成两类. 利用提出的基于模糊熵的层次聚类算法进行聚类, 结果如图 5 所示. 可以看出, 序列 7 和序列 8 被聚为一类, 称为 A 类, 其他 6 条序列为一类, 称为 B 类, 与几何图形的相似程度一致, 由此可知本文所提时间序列聚类算法结果与预期一致, 可较好区分不同类型的混沌时间序列.

其次, 选择 B 类序列 1 ∼ 6 进行预测. 由上述描述可知, 序列本身添加了白噪声, 因此可选择主曲线描述数据的分布特性. 此处采用基于主曲线的异常序列检测算法进一步剔除序列中的异常序列 (序

号 1, 2). 具体过程: 计算 B 类序列的均值, 并找到与其 DTW 距离最近的一条序列, 定义该序列的主曲线作为 B 类序列的主曲线, 如图 6 所示. 由图 6 可知, 主曲线能很好地反映数据分布的特性.

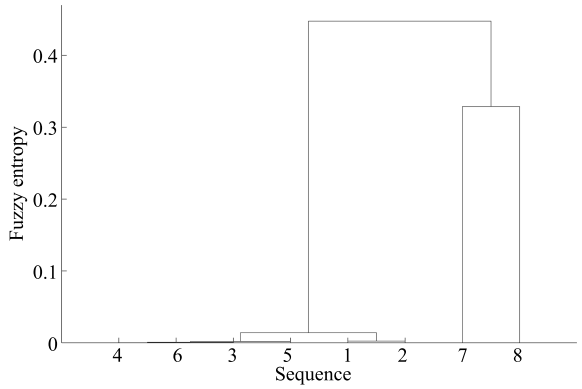


图 5 初始序列聚类结果

Fig. 5 The result of clustering on original sequences

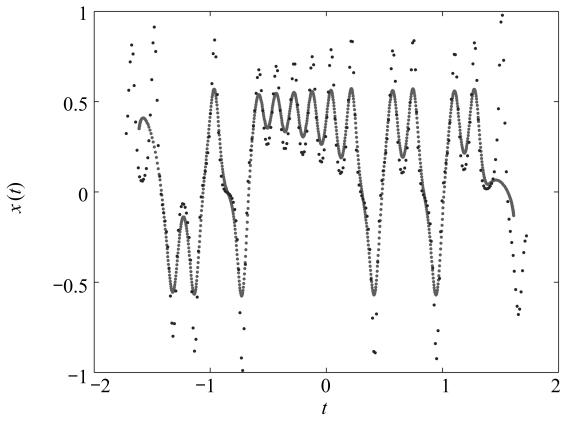


图 6 B 类序列的主曲线

Fig. 6 The principal curve of B class

再次, 利用式 (5) 计算 B 类各序列基于主曲线的异常因子. 由图 7 可以看出, 序列 1 和序列 2 的异常因子明显高于其他四条序列, 序列 3～6 的异常因子基本趋于一致. 因此可将序列 1 和序列 2 从该类剔除, 与生成数据时的设置完全一致.

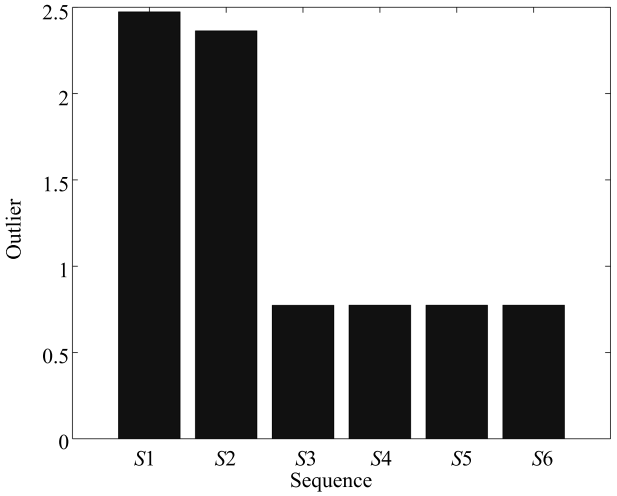


图 7 B 类各序列的异常因子

Fig. 7 Abnormal factor of every sequence in B class

为进一步分析本文所提方法的性能, 本文对初始序列、聚类阶段及异常序列剔除三个阶段分别进行预测. 限于篇幅, 以序列 3 为例, 预测效果如图 8 所示. 嵌入维 m 为 4, 时延 τ 为 1. 其中, 图 8 (a) 是针对初始 8 条序列同时回归建模的预测效果图, 图 8 (b) 是对已知序列进行聚类初步划分得到的 1～6 序列同时回归建模的预测效果图, 图 8 (c) 是在聚类初步划分后进行异常序列剔除挑选得到的序列 3～6 同时回归建模的预测效果图. 从图 8 可以看出, 随着相似序列的一步步筛选, 目标序列的预测效果明显变好, 同时验证了 M-SVR 建模效果取决于输出端的相似性, 越是相似的序列, 利用 M-SVR 进行预测效果越好.

表 1 给出了 3, 4, 5, 6 四条序列在三个阶段的预测效果. 由表 1 可知, 随着各个阶段序列相似程度的提高, 预测精度明显提高, 表明本文所提算法可以有效提高多变量时间序列的预测精度和数值稳定性, 进一步验证了 M-SVR 可有效利用多个输出端之间的结构化特性, 从而增加模型的信息含量, 提高建模的精度及稳定性.

以序列 3 为例, 图 9～11 分别给出本文所提方法, SVR 和 ELM 的预测效果及对应的局部放大图.

表 1 各个阶段预测结果性能指标对比

Table 1 Prediction performance parameters of capillary of three stages

序列	RMSE			MAE		
	初始序列	聚类后	异常序列剔除后	初始序列	聚类后	异常序列剔除后
序列 3	0.0589	0.0511	0.0319	0.0405	0.0357	0.0248
序列 4	0.0843	0.0814	0.0364	0.0638	0.0603	0.0270
序列 5	0.0559	0.0508	0.0379	0.0435	0.0387	0.0292
序列 6	0.0675	0.0585	0.0350	0.0494	0.0435	0.0269

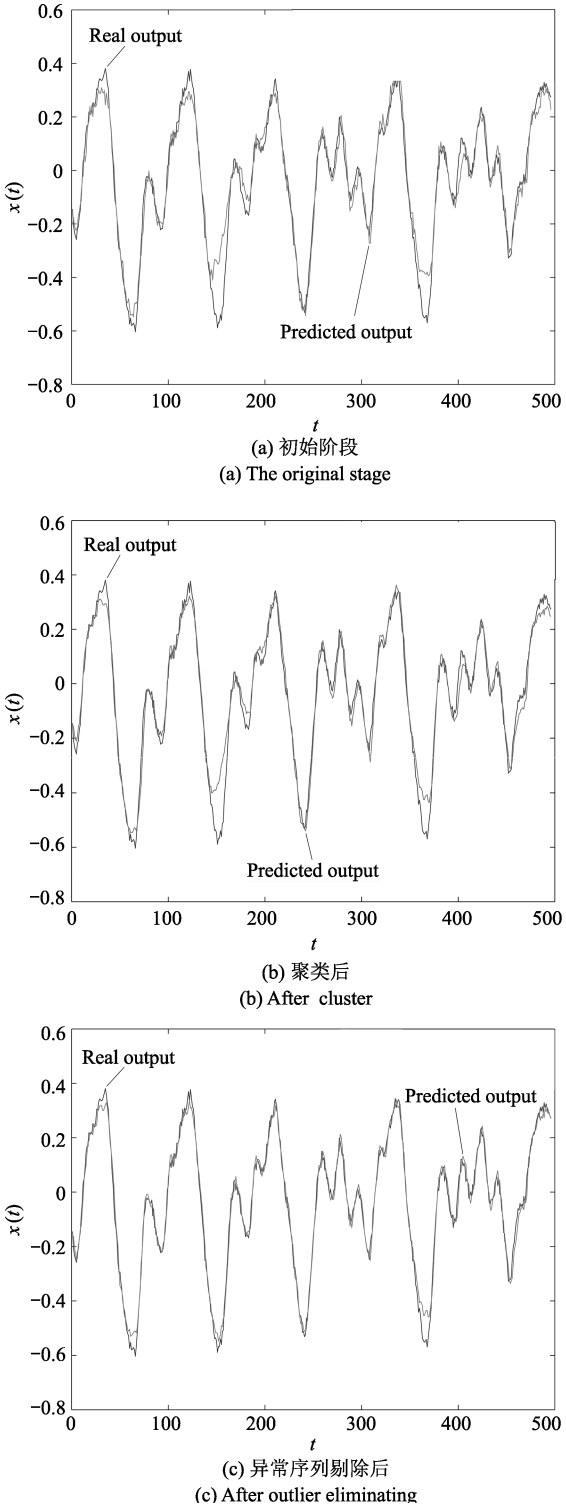


图 8　OE-MSVR 在三个阶段的预测效果

Fig. 8　The prediction of three stages with OE-MSVR algorithms

由图 9～11 可知, OE-MSVR 在整体上预测值与真实值最接近, 除了少数的波峰和波谷外, 大部分时间的预测曲线均贴近真实曲线; SVR 算法的预测值与真实值的走势基本相同, 但大部分时间点的预测值均与真实值存在偏差; ELM 算法的预测结果不太稳定, 相比其他两种算法与真实值的偏差相对较大.

从表 1 可以看出, 随着剔除相似度较低的序列, OE-MSVR 预测效果在不断提高, 表明异常序列的存在影响着数据集整体内在的关联性, 当筛选得到具有依赖关系较强相似性序列时, 结构化数据中蕴含的有价值的信息得到了有效利用. 结合图 9～11 可知, 在筛选得到的序列上进行操作时, OE-MSVR 预测效果明显好于其他两种算法, 进一步表明了 OE-MSVR 可以有效挖掘多变量时间序列的内在结构信息, 使得预测精度更高, 数值稳定性更好.

使用 OE-MSVR, SVR, ELM 对序列 3～6 等四条序列的预测效果对比如图 12 所示. 由图 12 可知, 本文所提算法在两种误差指标上明显比另外两种算法小, 预测效果明显较好, 再次表明本文方法对多变量时间序列的预测效果好, 验证了此方法的有效性与稳定性. 对于其他类别序列, 本文方法具有类似的对比效果.

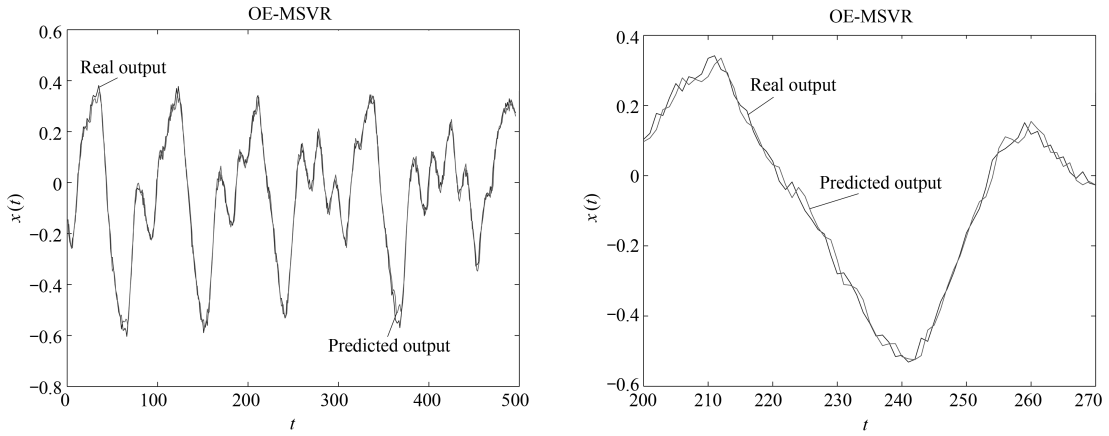
4.2　实际数据集时间序列预测

为验证算法的性能, 选择不同规模的数据集对算法进行测试, 并将本文所提算法与 SVM[22], ELM[21], Fol-BP[26] 及 AR[27] 算法进行对比. 其中, Fol-BP 是侯公羽等于 2014 年提出的多变量混沌时间序列预测算法, 通过 RMSE 和 MAE 进行评估. 实验设置与第 4.1 节相同. 为消除 ELM 和 Fol-BP 算法的随机性, 其结果为重复 10 次的平均值. 与第 4.1 节不同, 对于实际采集到的数据, 事先并没有其内在结构的先验知识.

实验中, OE-MSVR 均使用 RBF 核函数, 在澳门气象数据数据集上的核参数设为 2, 正则化参数 C 为 2^{-1}, 松弛变量 ε 为 0.01; 在 A monitor system 数据集上的核参数设为 2^3, 正则化参数 C 为 2^5, 松弛变量 ε 为 0.01; 在 Italian air quality 数据集上的核参数设为 2, 正则化参数 C 为 2^5, 松弛变量 ε 为 0.01; 在 Istanbul stock exchange 数据集上的核参数设为 1, 正则化参数 C 为 2^5, 松弛变量 ε 为 0.01; 在 Gas sensor array drift 数据集上的核参数设为 2^2, 正则化参数 C 为 2^2, 松弛变量 ε 为 0.01; Fol-BP, LibSVM 和 ELM 参数均为网络搜索得到最优值. 实验开始前, 所有的样本均归一化至 $[-1, 1]$.

4.2.1　数据集

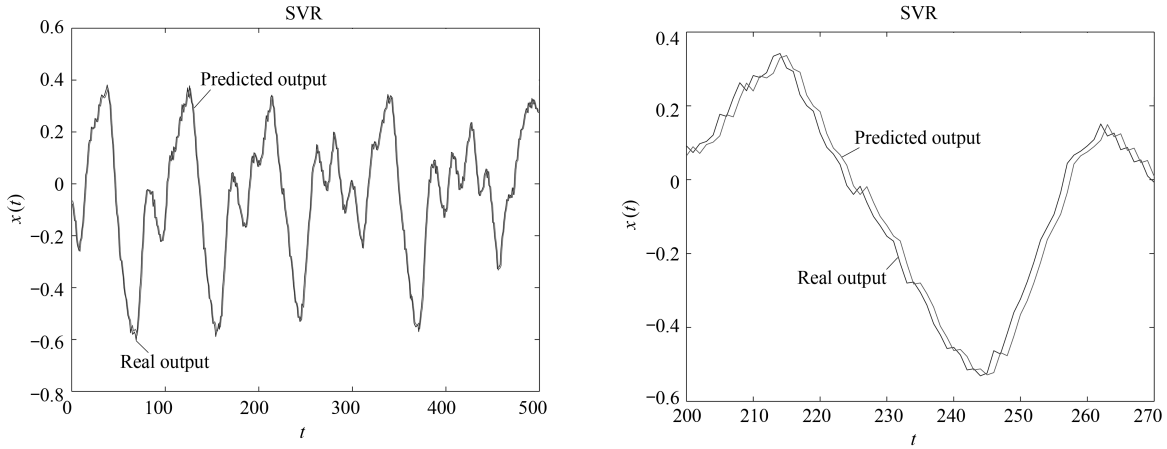
选定澳门气象数据, A monitor system, Italian air quality, Istanbul stock exchange 和 Gas sensor array drift 五种真实的时间序列数据集对 OE-MSVR 算法的性能进行验证. 使用的数据集可在澳门气象局官网、UCI 公共数据集下载得到. 五个真实数据集信息见表 2.



(a) 预测结果 (a) The prediction (b) 局部放大图 (b) The partially enlarged

图 9 序列 3 的 OE-MSVR 预测效果图

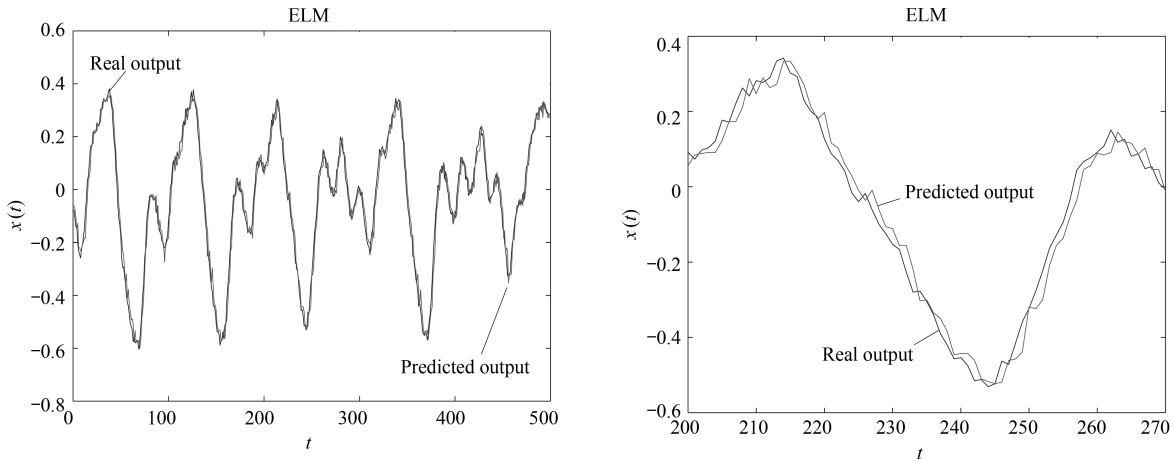
Fig. 9 The prediction of the third sequence with OE-MSVR algorithm



(a) 预测结果 (a) The prediction (b) 局部放大图 (b) The partially enlarged

图 10 序列 3 的 SVR 预测效果图

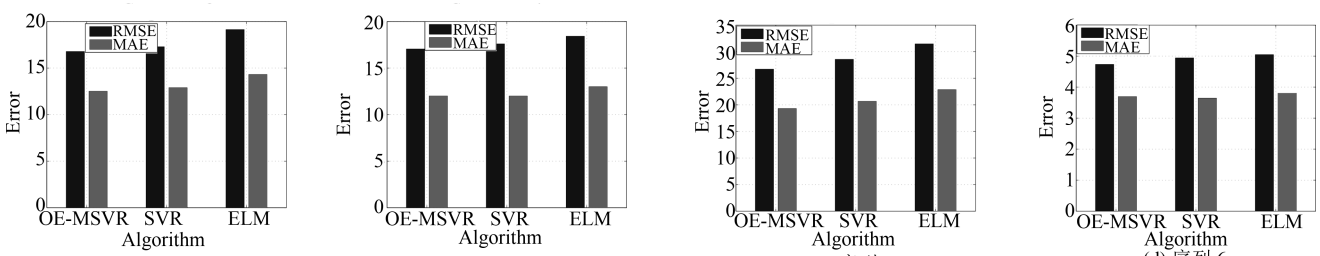
Fig. 10 The prediction of the third sequence with SVR algorithm



(a) 预测结果 (a) The prediction (b) 局部放大图 (b) The partially enlarged

图 11 序列 3 的 ELM 预测效果图

Fig. 11 The prediction of the third sequence with ELM algorithm



(a) 序列 3 (a) The third sequence　(b) 序列 4 (b) The fourth sequence　(c) 序列 5 (c) The fifth sequence　(d) 序列 6 (d) The sixth sequence

图 12　三种方法下预测误差对比图

Fig. 12　The errors of three algorithms

表 2　实际数据集信息

Table 2　Real datasets

数据集	样本数目		属性数目
	训练样本	测试样本	
澳门气象数据	1 276	547	11
Monitor system	1 260	540	19
Istanbul stock exchange	375	161	9
Air quality	1 680	720	13
Gas sensor array drift	310	133	30

本文首先选择澳门气象局官网提供的空气质量数据集做进一步对比实验. 该数据包括影响空气质量的多种因素: PM10, SO_2, NO_2, O_3, 伽马射线 (γ 射线), 气压 (P), 气温 (T), 湿度 (H), 风速 (W), 雨量 (R), 日照量 (S), 因此是一种典型的多变量时间序列数据. 对这些指标的同时预测, 具有明确的工程需求. 具体为 2002 年 ~ 2006 年 1 823 组大谭山的气象数据.

A monitor system, Italian air quality, Istanbul stock exchange 和 Gas sensor array drift 数据集均来源于 UCI 公共数据集. 其中 A monitor system 数据是安装在多人室中的监视器系统收集的数据, 包括 19 个连续属性; Italian air quality 数据是在意大利部署的气体多传感器收集到的空气质量数据, 包括 13 个空气质量属性; Istanbul stock exchange 数据集包括伊斯坦布尔证券交易所与七个其他国际指数的回报; Gas sensor array drift 数据是描述在不同浓度情况下气体传感器的漂移数据, 包括 30 个连续属性. 它们均属于典型的多变量时间序列数据.

4.2.2　实验结果

以澳门气象数据为例, 采用基于模糊熵的层次聚类对初始序列进行初步划分, 其结果如图 13 所示.

从图 13 可以看出, 聚类将序列 1, 3, 4, 7, 8, 9 聚为一类, 称作 A 类; 序列 2, 5, 6, 10 聚为一类, 称为 B 类; 序列 11 单独为一类, 称作 C 类.

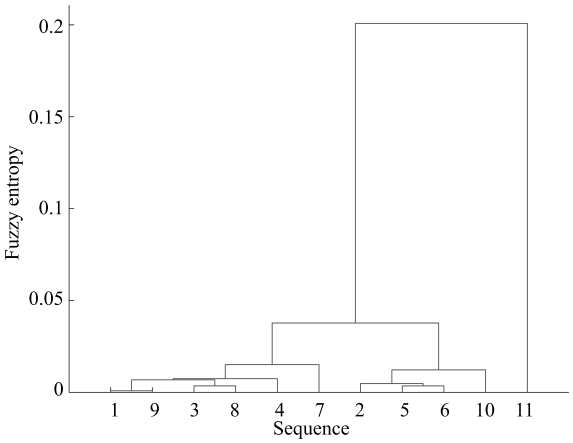


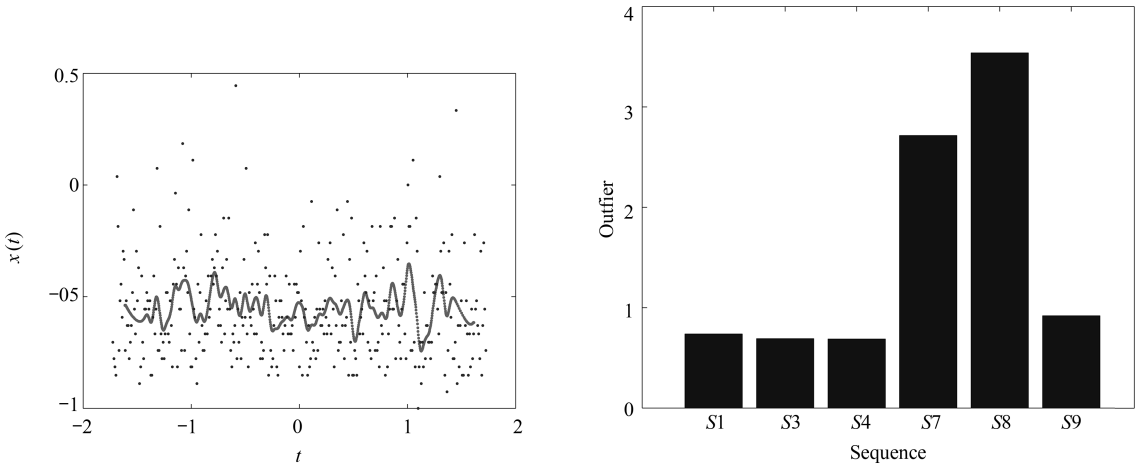
图 13　初始序列聚类结果

Fig. 13　The result of clustering on original sequences

以 A 类为例, 为得到相似度更高的序列, 计算各序列的异常因子. 构建 A 类序列的主曲线作为各序列异常因子检测的基准, 如图 14 (a) 所示; 并利用式 (5) 计算 A 类各个序列基于主曲线的异常因子, 如图 14 (b) 所示. 由图 14 可知, 序列 7 和序列 8 的异常因子明显较大, 故而将序列 7 和序列 8 从该类剔除.

利用筛选得到的序列构建模型, 用前 1 276 组数据训练, 后 547 组数据进行测试验证. 图 15 给出了五种算法的预测效果对比. 由图 15 可知, 虽然所提算法在较少序列上存在欠缺, 但在整体上 OE-MSVR 所提算法预测效果最好, 表明在处理实际多变量预测的问题时, OE-MSVR 有效提高了多条序列的预测精度, 进一步验证了所提方法对多变量时间序列预测的有效性与稳定性, 对实际多变量预测更具有实际的应用价值. 对于 B 类序列也有类似效果.

针对 A monitor system, Italian air quality, Istanbul stock exchange 和 Gas sensor array drift 四个 UCI 实际数据集, 其初始序列集在聚类阶段的结果如图 16 所示. 由图 16 可知, A monitor system 数据集中当聚类数目设置为 4 类时, 其中序列 1, 2,

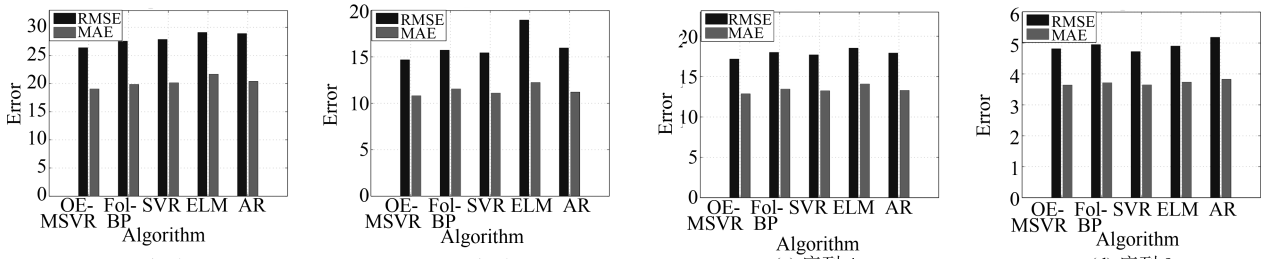


(a) A 类序列主曲线
(a) The principal curve of class A

(b) A 类各序列异常因子大小
(b) Abnormal factor of every sequence in class A

图 14 异常序列检测过程图

Fig. 14 The detection of abnormal sequences



(a) 序列 1
(a) The first sequence

(b) 序列 3
(b) The third sequence

(c) 序列 4
(c) The fourth sequence

(d) 序列 9
(d) The ninth sequence

图 15 五种方法下预测误差对比图

Fig. 15 The errors of five algorithms

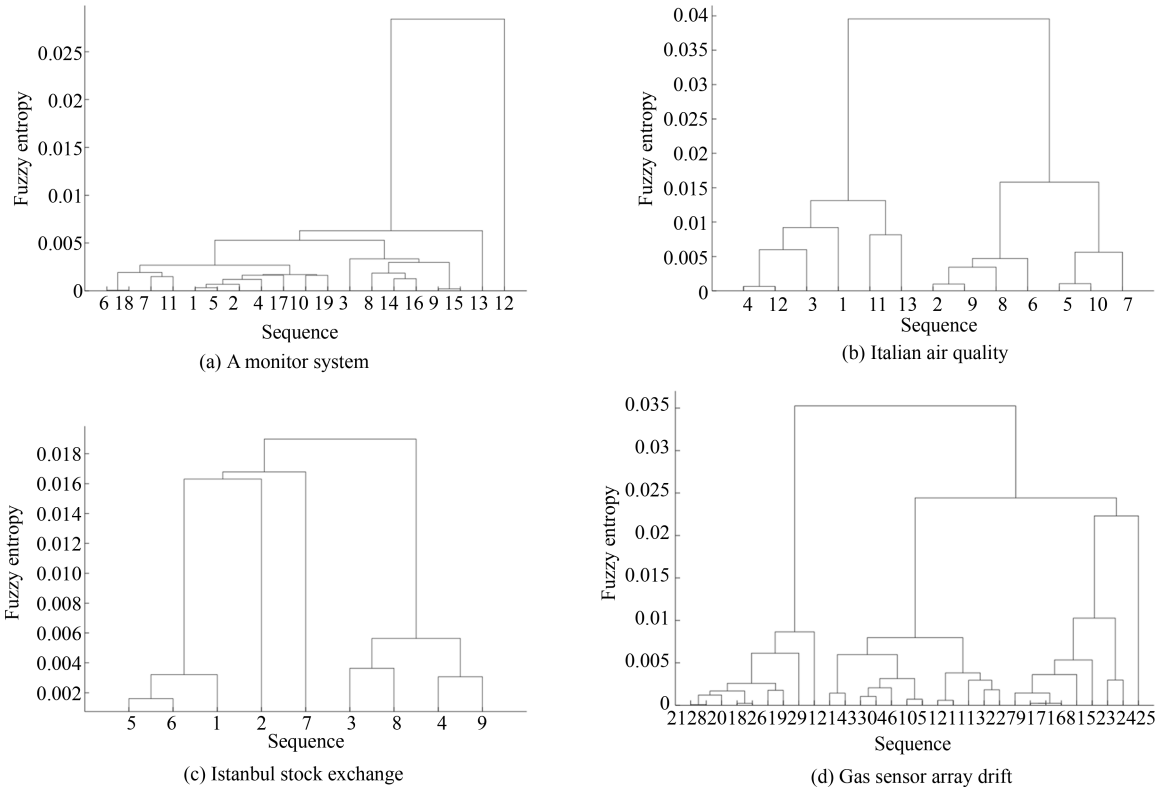
4, 5, 6, 7, 10, 11, 17, 18, 19 为一类, 称作 A 类; 序列 3, 8, 9, 14, 15, 16 为一类, 称作 B 类; 序列 12 和序列 13 分别为两类, 成为 C 类和 D 类. Italian air quality 数据集中当聚类数目设置为 2 类时, 其中序列 1, 3, 4, 11, 12, 13 为一类, 称作 A 类; 序列 2, 5, 6, 7, 8, 9, 10 为一类, 称作 B 类. Istanbul stock exchange 数据集中当聚类数目设置为 2 类时, 其中序列 1, 2, 5, 6, 7 为一类, 称作 A 类; 序列 3, 4, 8, 9 为一类, 称作 B 类; Gas sensor array drift 数据集中当聚类数目设置为 4 类时, 其中序列 1, 18, 19, 20, 21, 26, 28, 29 为一类, 称作 A 类; 序列 2, 3, 4, 5, 6, 10, 11, 12, 13, 14, 22, 27, 30 为一类, 称作 B 类; 序列 7, 8, 9, 15, 16, 17, 23, 24 为一类, 称作 C 类, 25 为一类, 称作 D 类.

针对各数据集的初始聚类结果, 本文以 A monitor system 数据集的 A 类, Italian air quality 数据集的 B 类, Istanbul stock exchange 数据集的 A 类、Gas sensor array drift 数据集的 B 类为例进行操作. 通过构建各类序列的主曲线以衡量序列集内部各序列的异常因子, 从而选取相似性高的序列集. 图 17 为四个数据集相应类的主曲线.

利用式 (5) 得到各序列基于主曲线的异常因子. 从图 18 可以看出, A monitor system, Italian air quality 和 Gas sensor array drift 数据集的各序列异常因子相对明显, Istanbul stock exchange 数据的异常因子相对不明显.

完成对各个数据集中相似序列的筛选后, 利用选择得到的序列集进行回归建模. 选择各数据集的前 70 % 组数据进行训练, 后 30 % 组数据测试验证, 具体样本数见表 2. 其中, OE-MSVR, Fol-BPELM 和 SVR 四种方法的嵌入维 m 为 2, 时延 τ 为 1, AR 算法的阶数为 1. 表 3 ∼ 6 分别给出了 A monitor system, Italian air quality, Istanbul stock exchange 和 Gas sensor array drift 四个实际数据集在 5 种算法下的预测效果对比.

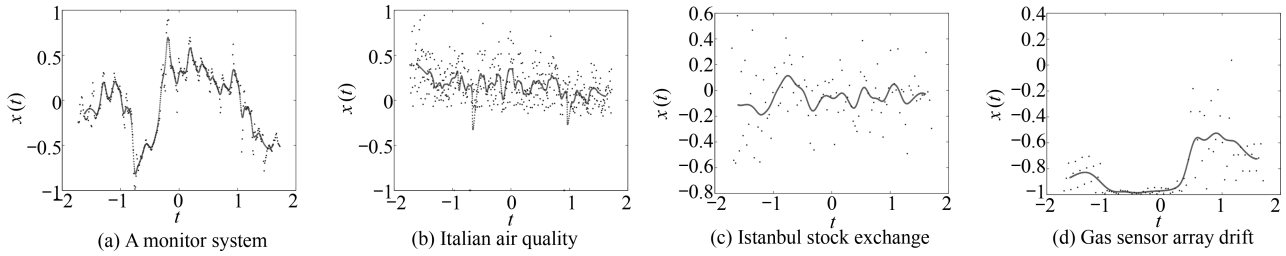
从表 3、表 4 和表 6 可以看出, 在大多数序列上, OE-MSVR 均明显优于其他对比算法, 尤其在表 3 的序列 1, 2; 表 4 的序列 2, 9, 10 和表 6 的序列 4, 5, 12, 13, 14 上, 本文所提方法均取得了较低的 RMSE 预测误差. 不难发现, Fol-BP 算法在数据



(a) A monitor system　(b) Italian air quality　(c) Istanbul stock exchange　(d) Gas sensor array drift

图 16　四个数据集的聚类结果图

Fig. 16　The results of clustering on four datasets



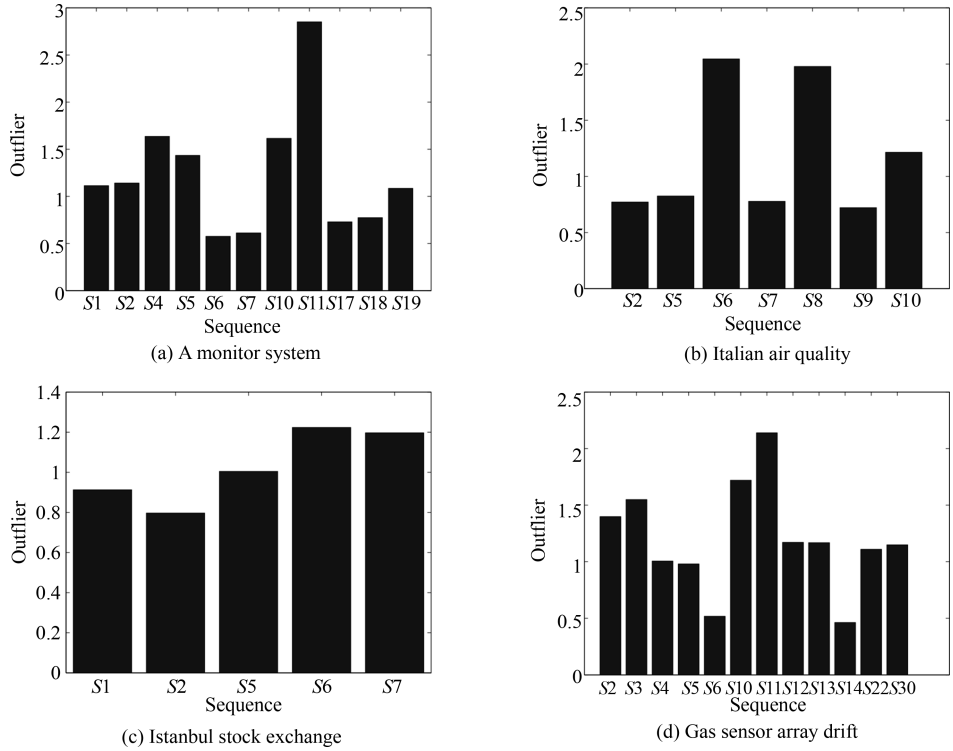
(a) A monitor system　(b) Italian air quality　(c) Istanbul stock exchange　(d) Gas sensor array drift

图 17　四个数据集对应类的主曲线结果图

Fig. 17　The principal curve of classes on four datasets

集 Italian air quality 上取得了较好的整体预测效果, 在序列 2, 7, 9, 10 上的 RMSE 预测误差均小于 ELM, SVR 和 AR, 表明较好利用多变量序列之间的结构化信息有助于降低整体的预测误差, 但由于该算法仅采用一阶局域法对多变量序列进行建模, 并没有深入考虑其间的结构化信息, 因此预测误差低于 OE-MSVR. 我们留意到 AR 在表 3 的序列 5, 19 等序列取得了最低的 RMSE 误差, 根据这些序列的图形走势可以看到, 该序列较为平缓, 波动不剧烈, AR 易于取得较好的预测效果, 而在序列波动相对较剧烈的表 4 中, AR 的预测误差则要明显高于本文所提算法和 Fol-BP, 也略高于 ELM, 与文献 [27] 的观测结果一致. 同时观察到, ELM 在少数序列上 (例如表 3 中序列 6, 19 和表 4 中序列 5) 的 RMSE 预测误差低于 OE-MSVR, 但这种结果来自于大量的网格搜索后取的最优值, 同时 ELM 本身也带有多输出的网络结构, 相比较而言, 本文所提算法参数为直接指定, 未做模型选择. SVR 的预测效果在所有算法中相对较差, 尽管其采用了网格搜索, 但由于 SVR 本身只能做单输出的预测, 因此在大多数序列上预测误差较高, 从另一个方面验证了多变量时间序列预测的必要性.

表 5 的结果验证了异常序列剔除的作用. 由图 18 (c) 可知, 与澳门气象数据 (图 14 (b)) 和 Italian air quality 数据 (图 18 (b)) 的异常因子分析对比, Istanbul stock exchange 数据集各序列异常因子区别并不显著. 当剔除具有相对较高异常因子的序列 6 和序列 7 后, OE-MSVR 虽然取得了最低的预测误



(a) A monitor system (b) Italian air quality

(c) Istanbul stock exchange (d) Gas sensor array drift

图 18 四个数据集的异常因子结果图

Fig. 18 The abnormal factors of four datasets

表 3 A monitor system 数据集预测结果性能指标对比

Table 3 Prediction performance parameters of capillary of A monitor system dataset

序列	RMSE					MAE				
	OE-MSVR	Fol-BP	SVR	ELM	AR	OE-MSVR	Fol-BP	SVR	ELM	AR
序列 1	0.1318	0.4561	0.8583	0.2388	0.1628	0.0949	0.3288	0.4115	0.1115	0.1628
序列 2	0.1006	0.5843	0.4862	0.1926	0.1514	0.0835	0.4236	0.3271	0.1249	0.1514
序列 5	4.7365	9.1443	14.5330	6.8052	4.0215	3.0607	6.4887	14.2472	2.4689	4.0215
序列 6	0.5226	1.6900	0.7693	0.4723	0.4642	0.3459	1.1693	0.6508	0.3407	0.4642
序列 7	0.3767	1.2893	0.5828	0.3817	0.2548	0.2730	0.9145	0.4679	0.2784	0.2548
序列 17	0.2112	0.9896	1.6971	0.4789	0.2820	0.1642	0.7497	0.7099	0.2239	0.2820
序列 18	0.7122	2.2701	0.9597	0.8947	0.6812	0.5146	1.7620	0.7590	0.6643	0.6812
序列 19	0.1244	0.7442	0.2993	0.0434	0.0994	0.0613	0.4913	0.2992	0.0038	0.0994

表 4 Italian air quality 数据集预测结果性能指标对比

Table 4 Prediction performance parameters of capillary of Italian air quality dataset

序列	RMSE					MAE				
	OE-MSVR	Fol-BP	SVR	ELM	AR	OE-MSVR	Fol-BP	SVR	ELM	AR
序列 2	121.9587	122.2608	123.6130	122.9571	123.4309	75.7143	70.2002	75.4696	76.6154	69.6428
序列 5	146.4420	148.3487	146.5258	146.3931	168.6011	89.7674	89.3968	88.6896	92.8062	100.1906
序列 7	113.6631	113.6741	117.6938	122.5415	123.6603	76.7395	77.8434	81.9074	85.8995	85.7884
序列 9	165.8359	167.1297	167.6215	174.2383	184.5600	100.3676	100.1063	101.4094	106.4807	108.7986
序列 10	178.1549	180.5234	186.9799	190.4919	200.0233	120.5245	118.7809	125.6156	129.6792	131.0641

表 5　Istanbul stock exchange 数据集预测结果性能指标对比

Table 5　Prediction performance parameters of capillary of Istanbul stock exchange dataset

序列	RMSE					MAE				
	OE-MSVR	Fol-BP	SVR	ELM	AR	OE-MSVR	Fol-BP	SVR	ELM	AR
序列 1	0.0119	0.0192	0.0121	0.0133	0.0119	0.0090	0.0140	0.0092	0.0101	0.0092
序列 2	0.0151	0.0217	0.0153	0.0167	0.0151	0.0116	0.0167	0.0117	0.0128	0.0117
序列 5	0.0097	0.0170	0.0095	0.0110	0.0098	0.0072	0.0123	0.0073	0.0083	0.0073

表 6　Gas sensor array drift 数据集预测结果性能指标对比

Table 6　Prediction performance parameters of capillary of Gas sensor array drift dataset

序列	RMSE					MAE				
	OE-MSVR	Fol-BP	SVR	ELM	AR	OE-MSVR	Fol-BP	SVR	ELM	AR
序列 2	9.08E+04	2.58E+05	9.17E+04	1.17E+05	9.27E+04	5.20E+04	1.58E+05	5.73E+04	7.48E+04	5.30E+04
序列 4	22.7062	74.2977	27.8097	26.3881	25.4101	10.9343	39.8237	17.5229	17.1492	15.2092
序列 5	31.5850	111.9757	34.3149	39.0925	34.9793	16.1337	73.2789	20.6630	25.5340	21.5229
序列 6	45.5174	221.2905	52.0124	59.9806	52.5336	22.6310	118.2681	27.9837	36.8887	30.6452
序列 12	18.0458	80.8880	24.1219	28.6046	20.4600	9.0439	45.7426	16.7021	16.8084	12.8685
序列 13	26.9213	79.6147	26.4391	34.9546	30.0822	13.2553	49.2278	16.5452	24.2313	19.3481
序列 14	36.5061	241.8456	38.7079	49.0467	43.4256	15.8653	138.9438	20.7230	27.4320	24.8211
序列 22	3.2263	5.0164	3.4150	2.7369	2.3271	2.7023	2.9809	2.9791	1.8935	1.4541
序列 30	3.1137	6.3610	3.4056	2.7941	2.2838	2.5935	4.2371	3.0138	1.8682	1.3841

差, 但是这种提高并不显著. 经过大量的网格搜索, SVR 和 ELM 也取得了较好的预测效果. 综合表 3～5 的结果, 我们发现, 当异常因子存在较大差异时 (如图 14 (b) 和图 18 (b)), 剔除异常序列后预测误差有显著下降 (如图 15 和表 4 所示); 而当异常序列并不显著时 (如图 18 (c)), 所提算法的整体预测效果与经过模型选择后的 SVR 和 ELM 等方法相仿, 提高幅度并不明显, 而这恰恰表明了异常序列在多变量时间序列预测中的负面影响.

5　结论

基于结构化输出的多变量时间序列预测可通过挖掘变量间蕴含的领域信息同时提高多个变量序列的预测效果. 其中的关键问题在于如何提取变量间的依赖关系. 本文提出了一种基于异常序列剔除的多变量时间序列预测方法. 该方法利用基于模糊熵的层次聚类对时间序列进行初步划分, 提出了基于主曲线的异常序列检测算法, 进一步检测并剔除异常序列, 最终引入多输出 SVR 进行建模和预测, 同时在理论上证明了该算法的可行性与合理性, 最终利用混沌时间序列数据与实际数据集数据验证了算法的有效性. 下一步的工作将集中在算法的泛化性理论分析和不同类型的变量间结构特性的建模.

References

1 Schölkopf B B, Smola A J. *Learning with Kernels*. Cambridge, Britain: MIT Press, 2002, **3**: 2165−2176

2 Zhang Yong, Guan Wei. Predication of multivariable chaotic time series based on maximal Lyapunov exponent. *Acta Physica Sinica*, 2009, **58**(2): 756−763
(张勇, 关伟. 基于最大 Lyapunov 指数的多变量混沌时间序列预测. 物理学报, 2009, **58**(2): 756−763)

3 Sun B Q, Guo H F, Karimi H R, Ge Y J, Xiong S. Prediction of stock index futures prices based on fuzzy sets and multivariate fuzzy time series. *Neurocomputing*, 2015, **151**: 1528−1536

4 Han Min, Xu Mei-Ling, Ren Wei-Jie. Research on multivariate chaotic time series prediction using mRSM model. *Acta Automatica Sinica*, 2014, **40**(5): 822−829
(韩敏, 许美玲, 任伟杰. 多元混沌时间序列的相关状态机预测模型研究. 自动化学报, 2014, **40**(5): 822−829)

5 Wang X Y, Han M. Improved extreme learning machine for multivariate time series online sequential prediction. *Engineering Applications of Artificial Intelligence*, 2015, **40**: 28−36

6 Han M, Xu M L, Liu X X, Wang X Y. Online multivariate time series prediction using SCKF-γ ESN model. *Neurocomputing*, 2015, **147**: 315−323

7 Chen T T, Lee S J. A weighted LS-SVM based learning system for time series forecasting. *Information Sciences*, 2015, **299**: 99−116

8 Sanchez-Fernandez M, de-Prado-Cumplido M, Arenas-Garcia J, Perez-Cruz F. SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems. *IEEE Transactions on Signal Processing*, 2005, **52**(8): 2298−2307

9 Bao Y K, Xiong T, Hu Z Y. Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing*, 2014, **129**: 482−493

10 Han J W, Kamber M, Pei J [Author], Fan Ming, Meng Xiao-Feng [Translator]. *Data Mining Concepts and Techniques (3rd edition) (Computer Science Series)*. Beijing: China Machine Press, 2012. 297−301

(Han J W, Kamber M, Pei J [著], 范明, 孟小峰 [译]. 数据挖掘概念与技术 (第 3 版) (计算机科学丛书). 北京: 机械工业出版社, 2012. 297−301)

11 Han Zhong-Ming, Chen Ni, Le Jia-Jin, Duan Da-Gao, Sun Jian-Zhi. An efficient and effective clustering algorithm for time series of hot topics. *Chinese Journal of Computers*, 2012, **35**(11): 2337−2347
(韩忠明, 陈妮, 乐嘉锦, 段大高, 孙践知. 面向热点话题时间序列的有效聚类算法研究. 计算机学报, 2012, **35**(11): 2337−2347)

12 Lee H. The Euclidean distance degree of Fermat hypersurfaces. *Journal of Symbolic Computation*, 2017, **80**: 502−510

13 Hautamaki V, Nykanen P, Franti P. Time-series clustering by approximate prototypes. In: Proceedings of the 19th International Conference on Pattern Recognition. Tampa, FL, USA: IEEE, 2008. 1−4

14 Yang Yi-Ming, Pan Rong, Pan Jia-Lin, Yang Qiang, Li Lei. A comparative study on time series classification. *Chinese Journal of Computers*, 2007, **30**(8): 1259−1266
(杨一鸣, 潘嵘, 潘嘉林, 杨强, 李磊. 时间序列分类问题的算法比较. 计算机学报, 2007, **30**(8): 1259−1266

15 Zhang Jun-Ping, Wang Yu. An overview of principal curves. *Chinese Journal of Computers*, 2003, **26**(2): 129−146
(张军平, 王珏. 主曲线研究综述. 计算机学报, 2003, **26**(2): 129−146)

16 Mao Wen-Tao, Wang Jin-Wan, He Ling, Yuan Pei-Yan. Hybrid sampling extreme learning machine for sequential imbalanced data. *Journal of Computer Application*, 2015, **35**(8): 2221−2226
(毛文涛, 王金婉, 何玲, 袁培燕. 面向贯序不均衡数据的混合采样极限学习机. 计算机应用, 2015, **35**(8): 2221−2226)

17 Sun Ke-Hui, He Shao-Bo, Yin Lin-Zi, A Di-Li · Duo Li-Kun. Application of fuzzyen algorithm to the analysis of complexity of chaotic sequence. *Acta Physica Sinica*, 2012, **61**(13): 130507
(孙克辉, 贺少波, 尹林子, 阿地力 · 多力坤. 模糊熵算法在混沌序列复杂度分析中的应用. 物理学报, 2012, **61**(13): 130507)

18 Mao Wen-Tao, Zhao Sheng-Jie, Zhang Jun-Na. Multi-input-multi-output support vector machine based on principal curve. *Journal of Computer Application*, 2013, **33**(5): 1281−1284, 1293
(毛文涛, 赵胜杰, 张俊娜. 基于主曲线的多输入多输出支持向量机算法. 计算机应用, 2013, **33**(5): 1281−1284, 1293)

19 Yuan P Y, Ma H D, Fu H Y. Hotspot-entropy based data forwarding in opportunistic social networks. *Pervasive and Mobile Computing*, 2015, **16**: 136−154

20 Tang Li-Dong, Song Bao-Wei, Li Zheng, Zheng Ke. A fuzzy reliability evaluation method for sub-sample products based on information entropy theory. *Journal of Projectiles, Rockets, Missiles and Guidance*, 2005, **25**(S1): 214−216
(汤礼东, 宋保维, 李正, 郑珂. 基于信息熵理论的小子样模糊可靠性评定方法. 弹箭与制导学报, 2005, **25**(S1): 214−216)

21 Mao W T, Zhao S J, Mu X X, Wang H C. Multi-dimensional extreme learning machine. *Neurocomputing*, 2015, **149**: 160−170

22 Chang C C, Lin C J. LIBSVM: a library for support vector machines [Online], available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html. February 1, 2016

23 Xu Mei-Ling, Han Min. Factor echo state network for multivariate chaotic time series prediction. *Acta Automatica Sinica*, 2015, **41**(5): 1042−1046
(许美玲, 韩敏. 多元混沌时间序列的因子回声状态网络预测模型. 自动化学报, 2015, **41**(5): 1042−1046)

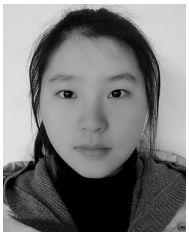
24 Li Jun, Li Da-Chao. Wind power time series prediction using optimized kernel extreme learning machine method. *Acta Physica Sinica*, 2016, **65**(13): 33−42
(李军, 李大超. 基于优化核极限学习机的风电功率时间序列预测. 物理学报, 2016, **65**(13): 33−42)

25 Ma Qian-Li, Zheng Qi-Lun, Peng Hong, Qin Jiang-Wei. Chaotic time series prediction based on fuzzy boundary modular neural networks. *Acta Physica Sinica*, 2009, **58**(3): 1410−1419
(马千里, 郑启伦, 彭宏, 覃姜维. 基于模糊边界模块化神经网络的混沌时间序列预测. 物理学报, 2009, **58**(3): 1410−1419)

26 Hou Gong-Yu, Liang Rong, Sun Lei, Liu Lin, Gong Yan-Fen. Risk analysis on long inclined-shaft construction in coalmine by TBM techniques based on multiple variables chaotic time series. *Acta Physica Sinica*, 2014, **63**(9): 90505
(侯公羽, 梁荣, 孙磊, 刘琳, 龚砚芬. 基于多变量混沌时间序列的煤矿斜井 TBM 施工动态风险预测. 物理学报, 2014, **63**(9): 90505)

27 Liu C H, Shang Y L, Duan L, Chen S P, Liu C C, Chen J. Optimizing workload category for adaptive workload prediction in service clouds. *Service-Oriented Computing. Lecture Notes in Computer Science*. Berlin, Heidelberg, Germany: Springer, 2015. 87−104

毛文涛 河南师范大学计算与信息工程学院副教授. 主要研究方向为机器学习, 时间序列预测. 本文通信作者.
E-mail: maowt@htu.edu.cn
(**MAO Wen-Tao** Associate professor at the College of Computer and Information Engineering, Henan Normal University. His research interest covers machine learning and prediction of time series. Corresponding author of this paper.)



蒋梦雪 河南师范大学计算机与信息工程学院硕士研究生. 主要研究方向为机器学习, 时间序列预测.
E-mail: jmxhtu@126.com
(**JIANG Meng-Xue** Master student at the College of Computer and Information Engineering, Henan Normal University. Her research interest covers machine learning and prediction of time series.)

李 源 河南师范大学计算与信息工程学院副教授. 主要研究方向为故障诊断, 可靠性预测.
E-mail: liyuan2015097@163.com
(**LI Yuan** Associate professor at the College of Computer and Information Engineering, Henan Normal University. Her research interest covers fault diagnosis and reliability prediction.)

张仕光 河南师范大学计算与信息工程学院副教授. 主要研究方向为机器学习, 大数据处理.
E-mail: 121114@htu.edu.cn
(**ZHANG Shi-Guang** Associate professor at the College of Computer and Information Engineering, Henan Normal University. His research interest covers machine learning and big data processing.)