(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110717543 B
(45)授权公告日 2023.09.19

(21)申请号 201910973966.6

(22)申请日 2019.10.14

(65)同一申请的已公布的文献号
申请公布号 CN 110717543 A

(43)申请公布日 2020.01.21

(73)专利权人 北京工业大学
地址 100124 北京市朝阳区平乐园100号

(72)发明人 乔俊飞 孙子健 汤健

(74)专利代理机构 北京思海天达知识产权代理有限公司 11203
专利代理师 刘萍

(51)Int.Cl.
G06F 18/21(2023.01)
G06F 18/22(2023.01)
G06F 18/2433(2023.01)
G06F 18/27(2023.01)
G06N 20/00(2019.01)

(56)对比文件
CN 108171251 A,2018.06.15
CN 106934035 A,2017.07.07
US 2003225525 A1,2003.12.04
US 2014122039 A1,2014.05.01
CN 107092582 A,2017.08.25
CN 105824785 A,2016.08.03
US 2019188212 A1,2019.06.20
CN 107402547 A,2017.11.28
CN 107403072 A,2017.11.28
CN 101827002 A,2010.09.08
韩法旺.数据流分类挖掘中的概念变化研究.《计算机科学》.2014,第41卷(第11期),第347-350+386页.
辛轶.A2-IKnnM-DHecoc:一种解决概念漂移问题的方法.《计算机研究与发展》.2011,第48卷(第4期),第592-601页.
郭躬德.一种适应概念漂移数据流的分类算法.《山东大学学报(工学版)》.2012,第42卷(第4期),第1-7页.
朱群.一种基于双层窗口的概念漂移数据流分类算法.《自动化学报》.2011,第37卷(第9期),第1077-1084页.
Denis dos Reis.Fast Unsupervised Online Drift Detection Using Incremental Kolmogorov-Smirnov Test.《KDD'16: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING》.2016,第1545-1554页.

审查员 刘书玲

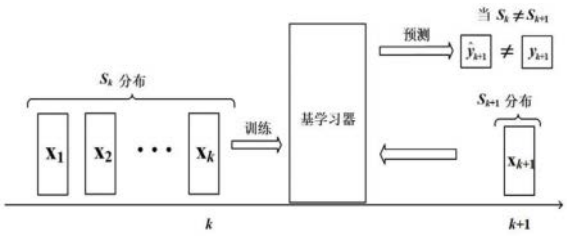
权利要求书2页 说明书10页 附图9页

(54)发明名称

基于样本分布统计检验的双窗口概念漂移检测方法

(57)摘要

基于样本分布统计检验的双窗口概念漂移检测方法属于机器学习领域。针对数据流随时间属性所具有的概念漂移问题,该方法首先在固定窗口内采用支持向量回归(SVR)进行离群点检测;然后针对检测到的离群点,在可变窗口内计算新旧样本间欧式距离,依据欧式距离,结合多种分布检验方法进行统计分析,以间接反映数据分布是否发生变化,进而确定是否发生漂移。最后在水泥强度基准数据集和城市固废焚烧(MSWI)出口氮氧化物浓度数据集上验证了所提方法的有效性。



CN 110717543 B

1.基于样本分布统计检验的双窗口概念漂移检测方法,其特征在于:

其中,X=[x_1,x_2,...,x_k]表示用于训练学习器的历史数据;学习器从样本x_{k+1}开始,随时间进行逐样本预测,当样本x_{k+w}被检测为离群点时,将目前w个样本与历史数据进行分布比对;w同时是可变窗口的大小;

第一个窗口用于离群点检测,目的是及时发现预测异常,以启动分布检测窗口;该窗口每次接收最新的一个样本,因此设置窗口大小为一个样本容量,窗口中使用SVR进行检测;

SVR其损失函数L_{loss}为:

$$L_{\text{loss}}(\hat{y}_i - y_i) = \begin{cases} 0, & |\hat{y}_i - y_i| < \varepsilon \\ |\hat{y}_i - y_i| - \varepsilon, & |\hat{y}_i - y_i| \geq \varepsilon \end{cases} \quad (i = 1,2,\ldots,k) \qquad (4)$$

其中,$\hat{y}_i$和y_i分别为训练集上的预测值和真实值,k为训练样本数量,ε是不敏感函数,代表了可接受误差的范围;

通过引入拉格朗日乘子α_i和$\hat{\alpha}_i$对回归问题进行优化,得到SVR函数:

$$\hat{y}_i = \sum_{i=1}^{k^*} (\hat{\alpha}_i - \alpha_i)\mathbf{x}_i^{\mathrm{T}}\mathbf{x} + b \qquad (5)$$

其中k^*为支持向量的个数,b为偏置:

$$b = y_i + \varepsilon - \sum_{i=1}^{k^*} (\hat{\alpha}_i - \alpha_i)\mathbf{x}_i^{\mathrm{T}}\mathbf{x} \qquad (6)$$

窗口内利用历史样本,建立SVR估计模型,对最新的数据进行回归估计,并依据误差e是否大于阈值ξ判断该样本是否为离群点,当$\lambda_{\text{test}}^{\text{outlier}} = 1$时向分布检测窗口发出报警;

误差阈值ξ计算公式如下:

ξ=(e_M-e_L)×0.4

其中,e_M为目前为止的最大预测误差,e_L为上一个样本的预测误差;

其中,

$$\lambda_{\text{test}}^{\text{outlier}} = \begin{cases} 1, & e > \xi \\ 0, & e < \xi \end{cases} \qquad (7)$$

当离群点检测窗口检测到异常并发出报警信号之后,分布检测窗口使用窗口内样本与历史样本进行匹配,判断报警是否来自概念漂移;在窗口内,首先计算历史样本中自身样本间的欧式距离,以及窗口内样本与历史样本间的欧式距离;在欧几里得空间中,样本x_i=[x_{i1},…,x_{iP}]和样本x_j=[x_{j1},…,x_{jP}](j=1,2,…,k)的欧式距离为:

$$d_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{iP} - x_{jP})^2} \qquad (8)$$

其中P表示每个样本的特征数量,s_{Old}^2和s_{New}^2分别表示原样本间距离D_{old}和新旧样本距离D_{New}的方差,

D_{old}=[d_{11},d_{12},…,d_{1k},…,d_{k1},d_{k2},…,d_{kk}] (9)

D_{New}=[d_{1k+1},d_{1k+2}…d_{1k+w}…$d_{kk+1}$$d_{kk+2}$$d_{kk+w}$] (10)

接着利用F检验，分析两组距离的方差齐性：

$$f^{\text{F-test}}(s_{\text{Old}}^2,\ s_{\text{New}}^2,\ F_\alpha^{\text{test}})=\frac{s_{\text{Old}}^2}{s_{\text{New}}^2} \tag{11}$$

F检验返回值$\lambda_{\text{test}}^{\text{F}}$为

$$\lambda_{\text{test}}^{\text{F}}=\begin{cases}0, & \text{在置信水平}F_\alpha^{\text{test}}\text{下方差相同}\\ 1, & \text{在置信水平}F_\alpha^{\text{test}}\text{下方差不相同}\end{cases} \tag{12}$$

当$\lambda_{\text{test}}^{\text{F}}=0$，两距离样本方差均为$s_p^2$时，利用t检验，对两组距离的平均值做进一步分析：

$$f^{\text{t-test}}(\overline{D_{\text{Old}}},\overline{D_{\text{New}}},\ \mu_{\text{ON}},\ S_{\text{Old}},\ S_{\text{New}},t_\alpha^{\text{test}})=\frac{\overline{D_{\text{Old}}}-\overline{D_{\text{New}}}-\mu_{\text{ON}}}{\sqrt{s_p^2/S_{\text{Old}}+s_p^2/S_{\text{New}}}} \tag{13}$$

其中μ_{ON}为两组距离样本总体的平均值之差，S_{Old}和S_{New}分别为两组距离样本的元素数量；t检验返回值$\lambda_{\text{test}}^{\text{t}}$为：

$$\lambda_{\text{test}}^{\text{t}}=\begin{cases}0, & \text{在置信水平}t_\alpha^{\text{test}}\text{下均数相同}\\ 1, & \text{在置信水平}t_\alpha^{\text{test}}\text{下均数不相同}\end{cases} \tag{14}$$

当$\lambda_{\text{test}}^{\text{t}}=0$时，两组距离具有相同分布，否则认为分布不同；

对于方差不相同的两组距离，采用曼-惠特尼U检验进行判断，其检验返回值$\lambda_{\text{test}}^{\text{U}}$为：

$$f^{\text{U-test}}(S_{\text{Old}},Z_{\text{Old}},U_\alpha^{\text{test}})=S_{\text{Old}}{}^2+S_{\text{Old}}(S_{\text{Old}}+1)/2-Z_{\text{Old}} \tag{15}$$

$$f^{\text{U-test}}(S_{\text{New}},Z_{\text{New}},U_\alpha^{\text{test}})=S_{\text{New}}{}^2+S_{\text{New}}(S_{\text{New}}+1)/2-Z_{\text{New}} \tag{16}$$

$$\lambda_{\text{test}}^{\text{U}}=\begin{cases}0, & \text{在置信水平}U_\alpha^{\text{test}}\text{下秩和相同}\\ 1, & \text{在置信水平}U_\alpha^{\text{test}}\text{下秩和不相同}\end{cases} \tag{17}$$

当$\lambda_{\text{test}}^{\text{U}}=0$时，两组距离具有相同分布，否则认为分布不同；则分布发生变化的条件为：

$$\begin{cases}\lambda_{\text{test}}^{\text{F}}=0\\ \lambda_{\text{test}}^{\text{t}}=1\end{cases}\ \text{or}\ \begin{cases}\lambda_{\text{test}}^{\text{F}}=1\\ \lambda_{\text{test}}^{\text{U}}=1\end{cases} \tag{18}$$

当发生概念漂移时，数据分布产生变化，这样的变化会反映在新旧数据间样本距离的变化上；两组距离在统计特性上存在的显著性差异，可间接反映数据的概念变化；如果在第二个窗口中未能检测到分布差异，则认为报警信号是由噪声引起，以此避免错误更新学习器。

基于样本分布统计检验的双窗口概念漂移检测方法

技术领域

[0001] 基于样本分布统计检验的双窗口概念漂移检测方法属于机器学习领域。

背景技术

[0002] 目前,机器学习的研究工作主要集中在非增量的批次学习上,其学习方法是将采集好的数据按批次打包为数据集合,集中训练基学习器。随着数据的海量增长,利用传统的数据集形式读取和处理数据使得数据的存储成本不断增加,同时集中训练的方式使得数据存在滞后性,无法及时反映当前时间工作状况,也不能对数据随时间的变化情况进行合理反馈。而在线学习算法基于单个样本或批量样本进行学习器的更新,进而期望得到基于目前全部样本的假设,该方式更适合实际问题。

[0003] 但事实上,在线学习过程中,数据并不始终满足同一分布,因此,数据的统计特性也随时间以某种方式产生变化,基于历史数据所得到的预测经验可能并不适用于当前决策,这将导致学习器的预测准确率下降。这种随时间推移,预测结果无法与真实数据匹配的现象称为概念漂移,本质上是数据的统计特性随时间发生了改变。

[0004] 针对概念漂移的检测,已有研究包括三种学习策略,分别为样本选择、样本加权和多概念集成学习。最常见的处理方法是样本选择,多采用窗口法,即对最新到达的一些样本在窗口中进行分析,以观察新样本与旧数据之间是否存在差异。

[0005] 本文在SVR的误差检测基础上,提出了利用三种假设检验类型,通过样本间欧式距离的分布相似性,检验样本之间是否存在概念漂移的方法。通过设置固定窗口与可变窗口,将方法分不同模块实施。两个窗口具有不同的检测任务,无需对参数及窗口尺寸进行调节,各个窗口对新样本按序计算,可节约计算成本和获得更快的检测能力。

[0006] 针对现实过程,其特性会随着环境的变化和时间的推移而发生改变,这一现象也预示着数据中所蕴涵的概念发生变化。如客户对一件商品的购买兴趣会随着时间变化,网站的访问量也会根据时间段产生不同,MSWI过程出口氮氧化物的浓度会由于季节变化和过程状态的调整而具有可变性。这种变化被叫做概念漂移或数据漂移。

[0007] 准确的来说,对于给定一系列带有标签的,截至到k时刻的历史数据$X=[x_1, x_2, \ldots, x_k]$,一般的学习目标是建立一个基学习器,并由这些历史数据进行训练,以使得在下一时刻k+1的数据到来时,尽可能准确的预测输出,即通过x_{k+1}预测y_{k+1}。

[0008] 对于下一时刻的样本x_{k+1},可认为其来自总体假设S_{k+1},历史数据满足总体假设S_k。当新旧数据服从同一分布时,$S_k=S_{k+1}$,代表数据是稳定的,当$S_k \neq S_{k+1}$时,认为数据不服从同一分布,即数据不稳定,发生概念漂移。这个过程如图1所示,其中$\hat{y}_{k+1}$为预测值。

[0009] 假设检验是根据样本数据推断总体数量特征的一种方法,用来判断样本与样本、样本与总体之间的差异是由抽样误差或本质差别引起,抽样误差和本质差别在漂移现象中可分别理解为噪声和分布变化。假设检验的目的是排除抽样误差的影响,判断样本间差别在统计意义上是否成立。其原理是先对总体的特征作某种假设,通过抽样研究推断该假设应该被拒绝或接受。常见的检验假设类型包括F检验、t检验、秩和检验。

[0010] F检验又称联合假设检验、方差齐性检验，根据样本之间方差关联程度判断样本相关性，为其他检验方法提供方差依据。样本M和样本N的方差分别为s_{M}^2和s_{N}^2时，在置信水平F_α^{test}下，F统计结果$f^{\mathrm{F\text{-}test}}$为：

[0011] $$f^{\mathrm{F\text{-}test}}(s_{\mathrm{M}}^2,\ s_{\mathrm{N}}^2,\ F_\alpha^{\mathrm{test}})=\frac{s_{\mathrm{M}}^2}{s_{\mathrm{N}}^2} \tag{1}$$

[0012] t检验用于检验样本平均值的差异，通过t分布理论推断差异发生的概率，从而判定两个平均值差异的显著性，进而判断样本之间的相关性。当样本具有相同方差s_p^2时，在置信水平t_α^{test}下，其检验结果$f^{\mathrm{t\text{-}test}}$为：

[0013] $$f^{\mathrm{t\text{-}test}}(\overline{M},\overline{N},\ \mu_0,\ S_{\mathrm{M}},\ S_{\mathrm{N}},t_\alpha^{\mathrm{test}})=\frac{\overline{M}-\overline{N}-\mu_0}{\sqrt{s_p^2/S_{\mathrm{M}}+s_p^2/S_{\mathrm{N}}}} \tag{2}$$

[0014] 其中μ_0为两组样本总体平均值之差，S_{M}和S_{N}分别为两组样本的元素数量，$\overline{M}$和$\overline{N}$为样本平均值。

[0015] 秩和检验的常用方法之一是曼-惠特尼U检验，属于非参数检验方法，在总体方差未知时，利用样本对总体分布形态进行推断。在置信水平U_α^{test}下，其检验结果$f^{\mathrm{U\text{-}test}}$为：

[0016] $$f^{\mathrm{U\text{-}test}}(S_{\mathrm{M}},Z_{\mathrm{M}},U_\alpha^{\mathrm{test}})=S_{\mathrm{M}}{}^2+S_{\mathrm{M}}\left(S_{\mathrm{M}}+1\right)/2-Z_{\mathrm{M}} \tag{3}$$

[0017] 其中Z_{M}为样本M的秩和。

[0018] 分析三种检验结果中差异是否显著，应预先设置三种检验的置信水平F_α^{test}、t_α^{test}、U_α^{test}，置信水平代表假设检验可以接受的误差范围，当置信水平过小时，分布检验对概念变化敏感，样本间存在轻度差别时，难以通过检验，导致漂移误判，加重学习器更新负担；当置信水平过大时，分布检验对概念变化表现宽容，导致漂移漏判，预测效果降低。因此在统计检验中，通常选取置信水平$F_\alpha^{\mathrm{test}},\ t_\alpha^{\mathrm{test}},\ U_\alpha^{\mathrm{test}}=5\%$。

发明内容

[0019] 图3为本文的算法结构。其中，$X=[x_1,x_2,\ldots,x_k]$表示用于训练学习器的历史数据。学习器从样本x_{k+1}开始，随时间进行逐样本预测，当样本x_{k+w}被检测为离群点时，将目前w个样本与历史数据进行分布比对。w同时是可变窗口的大小，取决于第一个开始检测的样本到异常样本之间的样本容量。

[0020] 概念漂移容易与噪声数据混淆，因为两者都会使数据分析产生偏差，但噪声仍然来自与历史数据相同的假设总体。因此合理辨别噪声是处理概念漂移问题的首要任务。

[0021] 本文第一个窗口用于离群点检测，目的是及时发现预测异常，以启动分布检测窗口。该窗口每次接收最新的一个样本，因此设置窗口大小为一个样本容量，窗口中使用SVR进行检测。

[0022] SVR是建立在支持向量上的回归分析，具有泛化能力强，学习速度快的优点，其损失函数L_{loss}为：

[0023] $$L_{\text{loss}}(\hat{y}_i - y_i) = \begin{cases} 0, & |\hat{y}_i - y_i| < \varepsilon \\ |\hat{y}_i - y_i| - \varepsilon, & |\hat{y}_i - y_i| \geq \varepsilon \end{cases} \quad (i = 1, 2, \ldots, k) \qquad (4)$$

[0024] 其中，$\hat{y}_i$和y_i分别为训练集上的预测值和真实值，k为训练样本数量，ε是不敏感函数，代表了可接受误差的范围。通过引入拉格朗日乘子α_i和$\hat{\alpha}_i$对回归问题进行优化，得到SVR函数：

[0025] $$\hat{y}_i = \sum_{i=1}^{k^*} (\hat{\alpha}_i - \alpha_i) \mathbf{x}_i^{\mathrm{T}} \mathbf{x} + b \qquad (5)$$

[0026] 其中k^*为支持向量的个数，b为偏置：

[0027] $$b = y_i + \varepsilon - \sum_{i=1}^{k^*} (\hat{\alpha}_i - \alpha_i) \mathbf{x}_i^{\mathrm{T}} \mathbf{x} \qquad (6)$$

[0028] 窗口内利用历史样本，建立SVR估计模型，对最新的数据进行回归估计，并依据误差e是否大于阈值ξ判断该样本是否为离群点，当$\lambda_{\text{test}}^{\text{outlier}} = 1$时向分布检测窗口发出报警。

[0029] 误差阈值ξ计算：

[0030] $\xi = (e_M - e_L) \times 0.4$

[0031] 其中，e_M为目前为止的最大预测误差，e_L为上一个样本的预测误差。

[0032] 其中，

[0033] $$\lambda_{\text{test}}^{\text{outlier}} = \begin{cases} 1, & e > \xi \\ 0, & e < \xi \end{cases} \qquad (7)$$

[0034] 当阈值ξ过小时，离群点检测窗口会频繁报警，分布检测窗口长期处于检测状态，计算时间增加，检测效率下降；当阈值ξ过大时，离群点检测窗口报警频率下降，分布检测窗口可能错过开始发生漂移的样本，检测效率下降。因此，根据平稳时数据的波动状况选取合理的阈值，可增加预测精度，提高分布检测效率。

[0035] 当离群点检测窗口检测到异常并发出报警信号之后，分布检测窗口使用窗口内样本与历史样本进行匹配，判断报警是否来自概念漂移。在窗口内，首先计算历史样本中自身样本间的欧式距离，以及窗口内样本与历史样本间的欧式距离。在欧几里得空间中，样本$x_i = [x_{i1}, \cdots, x_{iP}]$和样本$x_j = [x_{j1}, \cdots, x_{jP}]$(j=1,2,…,k)的欧式距离为：

[0036] $$d_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{iP} - x_{jP})^2} \qquad (8)$$

[0037] 其中P表示每个样本的特征数量，s_{Old}^2和s_{New}^2分别表示原样本间距离D_{old}和新旧样本距离D_{New}的方差，

[0038] $D_{old} = [d_{11}, d_{12}, \cdots, d_{1k}, \cdots, d_{k1}, d_{k2}, \cdots, d_{kk}]$ (9)

[0039] $D_{New} = [d_{1k+1}, d_{1k+2} \cdots d_{1k+w} \cdots d_{kk+1} d_{kk+2} d_{kk+w}]$ (10)

[0040] 接着利用F检验，分析两组距离的方差齐性：

[0041] $$f^{\text{F-test}}(s_{\text{Old}}^2, s_{\text{New}}^2, F_\alpha^{\text{test}}) = \frac{s_{\text{Old}}^2}{s_{\text{New}}^2} \qquad (11)$$

[0042] F检验返回值$\lambda_{\text{test}}^{\text{F}}$为

[0043] $$\lambda_{\text{test}}^{\text{F}}=\begin{cases}0, & \text{在置信水平}F_{\alpha}^{\text{test}}\text{下方差相同}\\ 1, & \text{在置信水平}F_{\alpha}^{\text{test}}\text{下方差不相同}\end{cases} \tag{12}$$

[0044] 当$\lambda_{\text{test}}^{\text{F}}=0$，两距离样本方差均为$s_p^2$时，利用t检验，对两组距离的平均值做进一步分析：

[0045] $$f^{\text{t-test}}(\overline{D_{\text{Old}}},\overline{D_{\text{New}}},\ \mu_{\text{ON}},\ S_{\text{Old}},\ S_{\text{New}},t_{\alpha}^{\text{test}})=\frac{\overline{D_{\text{Old}}}-\overline{D_{\text{New}}}-\mu_{\text{ON}}}{\sqrt{s_p^2/S_{\text{Old}}+s_p^2/S_{\text{New}}}} \tag{13}$$

[0046] 其中μ_{ON}为两组距离样本总体的平均值之差，S_{Old}和S_{New}分别为两组距离样本的元素数量。t检验返回值$\lambda_{\text{test}}^{\text{t}}$为：

[0047] $$\lambda_{\text{test}}^{\text{t}}=\begin{cases}0, & \text{在置信水平}t_{\alpha}^{\text{test}}\text{下均数相同}\\ 1, & \text{在置信水平}t_{\alpha}^{\text{test}}\text{下均数不相同}\end{cases} \tag{14}$$

[0048] 当$\lambda_{\text{test}}^{\text{t}}=0$时，两组距离具有相同分布，否则认为分布不同。

[0049] 对于方差不相同的两组距离，采用曼-惠特尼U检验进行判断，其检验返回值$\lambda_{\text{test}}^{\text{U}}$为：

[0050] $$f^{\text{U-test}}(S_{\text{Old}},Z_{\text{Old}},U_{\alpha}^{\text{test}})=S_{\text{Old}}{}^2+S_{\text{Old}}(S_{\text{Old}}+1)/2-Z_{\text{Old}} \tag{15}$$

[0051] $$f^{\text{U-test}}(S_{\text{New}},Z_{\text{New}},U_{\alpha}^{\text{test}})=S_{\text{New}}{}^2+S_{\text{New}}(S_{\text{New}}+1)/2-Z_{\text{New}} \tag{16}$$

[0052] $$\lambda_{\text{test}}^{\text{U}}=\begin{cases}0, & \text{在置信水平}U_{\alpha}^{\text{test}}\text{下秩和相同}\\ 1, & \text{在置信水平}U_{\alpha}^{\text{test}}\text{下秩和不相同}\end{cases} \tag{17}$$

[0053] 当$\lambda_{\text{test}}^{\text{U}}=0$时，两组距离具有相同分布，否则认为分布不同。则分布发生变化的条件为：

[0054] $$\begin{cases}\lambda_{\text{test}}^{\text{F}}=0\\ \lambda_{\text{test}}^{\text{t}}=1\end{cases}\ \text{or}\ \begin{cases}\lambda_{\text{test}}^{\text{F}}=1\\ \lambda_{\text{test}}^{\text{U}}=1\end{cases} \tag{18}$$

[0055] 当发生概念漂移时，数据分布产生变化，这样的变化会反映在新旧数据间样本距离的变化上。两组距离在统计特性上存在的显著性差异，可间接反映数据的概念变化。如果在第二个窗口中未能检测到分布差异，则认为报警信号是由噪声引起，以此避免错误更新学习器。

[0056] 图4中描述了本方法的检验流程。新样本到达后，利用训练后的SVR对该样本测试，当测试误差小于阈值ξ时，认为数据没有产生变化，当测试误差大于阈值ξ时，认为数据异常。接着计算可变窗口内样本与历史样本之间的欧式距离，通过F检验观察两组距离数据在方差上是否有相似性。当方差无显著性差异时，对两组距离数据进行t检验，以两组距离平均值的相似性判断两组样本之间的相关性与分布情况。当方差存在显著性差异时，通过秩和检验，分析两组样本之间的漂移状况。

附图说明

[0057] 图1.概念漂移的一般形式

[0058] 图2.(a)交替型窗口(b)竞争型窗口(c)自适应大小窗口

[0059] 图3.算法结构

[0060] 图4.算法流程图

[0061] 图5.学习器在基准数据上训练效果

[0062] 图6.(a)测试集1的拟合效果(b)测试集1的拟合误差(c)测试集2的拟合效果(d)测试集2的拟合误差(e)测试集3的拟合效果(f)测试集3的拟合误差

[0063] 图7.(a)训练集样本之间的距离分布(b)训练集和测试集1之间的距离分布(c)训练集和测试集2之间的距离分布(d)训练集和测试集3之间的距离分布

[0064] 图8.学习器在工业数据上训练效果

[0065] 图9.(a)测试集1的拟合效果(b)测试集1的拟合误差(c)测试集2的拟合效果(d)测试集2的拟合误差(e)测试集3的拟合效果(f)测试集3的拟合误差

[0066] 图10.(a)训练集样本之间的距离分布(b)训练集和测试集1之间的距离分布(c)训练集和测试集2之间的距离分布(d)训练集和测试集3之间的距离分布

具体实施方式

[0067] 为验证本方法性能,本文选用水泥强度基准数据集进行测试,数据来自Prof.I-Cheng Yeh的学习团队,可通过访问UCI(https://archive.ics.uci.edu/)获取。数据集共包含样本1030个,输入变量是直接或间接影响水泥抗压强度的主要因素,分别是水泥(Cement)、高炉渣(Blast Furnace Slag)、粉煤灰(Fly Ash)、水(Water)、高效减水剂(Superplasticizer)、粗骨料(Coarse Aggregate)、细骨料(Fine Aggregate)、龄期(Age),输出为混凝土抗压强度(Concrete compressive strength)。

[0068] 首先将数据集划分为两个子数据集,分别包含原数据集中前500组和后500组数据,然后将两个子数据集等间隔分为五份,每份包含100组数据。本文使用第一个子数据集中的第1份数据作为训练集进行建模,使用第一个子数据集中的第3份数据、第4份数据和第二个子数据集的第1份数据分别作为测试集1、测试集2和测试集3进行测试。实验中,只对数据做标准化处理,不进行特征选择,并在测试时逐样本拟合,目的是模拟实时处理环境。

[0069] 图5展示了训练后的SVR对于训练集的拟合效果,在SVR中,核函数选择为RBF(Radial Basis Function),惩罚参数为1,核函数宽度σ为1,不敏感度μ=0.001。误差阈值ξ设置为25(阈值大小根据图6中的拟合误差,人为选定),即当拟合误差超过25时窗口发出报警。

[0070] 图6展示了在测试集1、测试集2和测试集3上的拟合效果与误差曲线。可看出三个测试集分别在第35个样本,第10个样本,第24个样本上出现第一个离群点,同时学习器对测试集1,3的拟合效果较为平稳,对测试集2的拟合效果波动较大。由于SVR的特性,学习器容易识别出具有较高预测误差的样本,表明本文第一个窗口在检测离群点的工作上表现灵敏并且精准。

[0071] 本文使用第二个窗口,分别计算训练集中各样本间距离、测试集1与训练集中各样本间距离、测试集2与训练集中各样本间距离、测试集3与训练集中各样本间距离。对以上4个距离集合进行F检验、t检验或曼-惠特尼U检验,并依次设置三种检验的显著性水平 F_{α}^{test}, t_{α}^{test}, $U_{\alpha}^{\text{test}}=1\%$, F_{α}^{test}, t_{α}^{test}, $U_{\alpha}^{\text{test}}=3\%$, F_{α}^{test}, t_{α}^{test}, $U_{\alpha}^{\text{test}}=5\%$。

[0072] 对于三种检验类型，h＝0代表接受原假设，即两组距离数据之间存在相同的方差或平均值，有强相关性；h＝1代表拒绝原假设，即两组距离数据之间不存在相同的方差或平均值，无强相关性。h的数值由各检验统计量值和其对应的临界值表所确定。表1～表3展示了对不同测试集在三种显著性水平下的检验结果。

[0073] 表1.假设检验结果(α＝0.01)

[0074]

F-α，t-α，U-α=1%	F-test	T-test	U-test
Test1	1	-	0
Test2	1	-	1
Test3	1	-	0

[0075] 表2.假设检验结果(α＝0.03)

[0076]

F-α，t-α，U-α=3%	F-test	T-test	U-test
Test1	0	0	-
Test2	1	-	1
Test3	1	-	0

[0077] 表3.假设检验结果(α＝0.05)

[0078]

F-α，t-α，U-α=5%	F-test	T-test	U-test
Test1	0	0	-
Test2	1	-	1
Test3	1	-	0

[0079] 根据检验结果，测试集1所对应的距离集合和训练样本自身距离具有相似的方差，通过了t检验。测试集2和测试集3对应的距离集合与训练样本自身距离具有不同的方差，但测试集2未能通过U检验。图7中给出各个距离集合所对应的分布图。

[0080] 表4展示了四组距离在不同区间内的分布情况。由此可见，训练集中样本距离集中在0～0.5区间内，而测试集2和训练集之间的样本距离在该区间内占比较小。这揭示了数据之间存在的概念漂移，验证了分布检测窗口的有效性。

[0081] 表4.基准数据集上不同距离区间的样本分布

[0082]

范围 ($\times10^5$)	0～0.5	0.5～1	1～1.5	1.5～2	2～2.5
Train	0.3925	0.2755	0.2243	0.0913	0.0164
Test1	0.2227	0.3558	0.2653	0.1218	0.0345
Test2	0.1400	0.3540	0.3490	0.1430	0.0140
Test3	0.1964	0.3679	0.2887	0.1066	0.0405

[0083] 本文在基准数据集上与基于熵的概念漂移检测方法进行比较。在两个数据集中，分别计算4组距离落在0～0.5,0.5～1,1～1.5,1.5～2,2～2.5范围内的比例，并计算熵值，同时用本文方法对各范围内数据进行统计检验分析，将结果记录在表5中。

[0084] 表5.在基准数据集上的方法比较

[0085]

范围($\times10^5$)		0～0.5	0.5～1	1～1.5	1.5～2	2～2.5	总体
		F-test,T-test,U-test(α=0.05)					
Test1	基于熵	0.0022	0.0054	0.0155	0.0165	0.0023	0.0084
	基于检验统计	0 0 -	0 0 -	0 0 -	0 0 -	0 1 -	0 0 -
Test2	基于熵	0.0049	0.0160	0.0282	0.0409	0.0238	0.0227
	基于检验统计	1 - 1	1 - 1	0 1 -	0 0 -	0 0 -	1 - 1
Test3	基于熵	0.0007	0.0105	0.0110	0.0332	0.0005	0.0112
	基于检验统计	0 0 -	0 0 -	0 0 -	0 0 -	0 1 -	1 - 0

[0086]

[0087] 基于熵的方法在数据分布相等时，熵值为1，分布不同时，熵值为0。在表5中，测试集2的平均熵值更接近1，认为分布未发生变化，而根据本文方法得到的距离分布图和统计检验结果，看出测试集2在0～1范围内未能通过秩和检验，认为分布发生变化，这一结果也与学习器在测试集2上的拟合结果相印证。因此基于熵的方法相比与统计检验算法在连续型变量的检验上表现欠佳。

[0088] 本文工业数据来自国内某MSWI发电厂。在MSWI的焚烧阶段会产生包括氮氧化物在内的大量烟气，为防止烟气形成二次污染，可通过对出口氮氧化物的浓度进行预测，并依据预测结果及时采取相应的防治手段。但受不同季节固废水分含量变化、焚烧炉内环境变化等因素影响，不同工况下的氮氧化物排出浓度其预测模型参数可能不相同，这一现象符合本文所要解决的概念漂移问题。

[0089] 本文考虑了氮氧化物的生产端与消除端，选取了与炉膛温度、一次风量、二次风

量、炉膛剩余氧量、尿素喷入量等因素相关性最强的19个变量，依据时间顺序选取1000个样本为训练集，另1500个样本等间隔划分为三个测试集。

[0090] 在训练学习器过程中，设置惩罚参数为20，核函数宽度σ为20，不敏感度μ＝0.001，三种假设检验的显著性水平设置为F_{α}^{test}，t_{α}^{test}，$U_{\alpha}^{\text{test}}=5\%$。验证效果如图8、图9。

[0091] 图9展示了在工业数据集上的预测和误差曲线，可以看出当测试集所处工况与训练集相同时，学习器预测精准，而当工况转变，预示概念发生变化时，预测结果出现较大误差。进一步对测试集与训练集在显著性水平α＝0.05进行假设检验测试：

[0092] 表6.假设检验结果(α＝0.05)

[0093]

F-α，t-α，U-α=5%	F-test	T-test	U-test
Test1	1	-	1
Test2	1	-	1
Test3	1	-	1

[0094] 从表6看出，由于三个测试集中工况变化较为明显，因此其距离分布相较于训练集中的距离分布方差不同，未能通过F检验，并且未能通过U检验也标志着数据中存在着分布的变化，因此存在概念漂移。

[0095] 图10和表7中可以看出三个测试集与训练集的距离分布差异，表现为在0～0.5距离范围内，训练集的样本占比较高，而三个测试集的距离图像在此区间较为低矮，有更多的样本距离落在0.5～1之间。这同样验证了在工业数据集中的概念漂移现象。

[0096] 表7.工业数据集上不同距离区间的样本分布

[0097]

范围 ($\times 10^4$)	0～0.5	0.5～1	1～1.5	1.5～2	2～2.5
Train	0.8437	0.1471	0.0390	0.0070	0.0018
Test1	0.7568	0.1848	0.0459	0.0124	0.0001
Test2	0.7907	0.1621	0.0344	0.0103	0.0025
Test3	0.7736	0.1599	0.0443	0.0166	0.0056

[0098] 本文在工业数据集上同样与基于熵的检测方法进行比较。在两个数据集中，分别计算4组距离落在0～0.5，0.5～1，1～1.5，1.5～2，2～2.5范围内的比例，并计算熵值，同时用本文方法对各范围内数据进行统计检验分析，将结果记录在表8中。

[0099] 表8.在工业数据集上的方法比较

[0100]

范围(×10⁴)		0～0.5	0.5～1	1～1.5	1.5～2	2～2.5	总体
		F-test,T-test,U-test(α=0.05)					
Test1	基于熵	0.0598	0.0510	0.0264	0.0298	0.0142	0.0362
	基于检验统计	1 - 1	1 -- 1	1 - 1	1 - 1	1 - 1	1 - 1
Test2	基于熵	0.0235	0.0263	0.0103	0.0192	0.0058	0.0170
	基于检验统计	1 - 1	1 - 1	1 - 1	1 - 1	1 - 0	1 - 1
Test3	基于熵	0.0424	0.0235	0.0217	0.0493	0.0263	0.0326
	基于检验统计	1 - 1	1 - 1	1 - 1	1 - 0	1 - 0	1 - 1

[0101] 针对表8中的测试集2,尽管两种检验方法的结果相同,但本文方法可额外反映出测试集中的方差变化,方差代表着数据的变异程度,这为后续学习器模型的更新提供重要依据。因此本文方法相比基于熵的检测方法可以在各区间上给出详细的统计信息,能观察到数据整体的变化幅度与变化形式。

[0102] 此处主要分析不同参数对检验结果的影响。对于阈值ξ:根据图6可知,其大小决定离群点检测窗口的报警频率,影响学习器的内存占用和计算能力,因此决定预测精度,进而影响分布检验效率。

[0103] 对于置信水平F_{α}^{test}、t_{α}^{test}、U_{α}^{test}:根据表1～表3可知,与阈值ξ的影响效果相似,其大小改变会使分布检验窗口对于数据变化的敏感度不同,影响分布检测的及时性。

[0104] 对于三种检验的返回值λ_{test}^{F}、λ_{test}^{t}、λ_{test}^{U}:由以上实验结果可得,当预测误差大幅上升时,数据之间距离的方差也会变为不相等,即$\lambda_{test}^{F}=1$;而当预测准确率持续大幅变化时,数据之间距离的均值和秩和等级也会不相等,即$\lambda_{test}^{t}=1$和$\lambda_{test}^{U}=1$。如在表5和图6中,测试集2上的预测效果大幅下降,F检验返回值$\lambda_{test}^{F}=1$,同时预测准确率频繁波动,其U检验返回值$\lambda_{test}^{U}=1$,揭示了测试集2中数据相较训练集的不同分布。测试集3上存在较大预测误差,F检验返回值$\lambda_{test}^{F}=1$,但其预测准确率较为平稳,U检验返回值$\lambda_{test}^{U}=0$。在表8和图9中,三个测试集上的预测误差与预测准确率均呈现较大波动,其都具有$\lambda_{test}^{F}=1$和$\lambda_{test}^{U}=1$。因此F检验返回值λ_{test}^{F}具有跟踪离群点的能力,可在分布可能变化时进行反馈。而t检验和U检验的返回值λ_{test}^{t}和λ_{test}^{U}具有跟踪分布状况的能力,可在F检验基础上确定分布是否变化。

[0105] 本文提出了一种基于SVR检测和基于欧式距离统计检验的双窗口在线概念漂移检测方法,并在水泥强度基准数据集和MSWI工业数据集上进行测试。本文的主要贡献是:(1)本文提出基于双窗口的概念漂移检测,先基于第一个窗口进行异常检测,再基于第二个窗

口采用3种假设检验方式进行实时的数据分布检测;(2)本文提出通过样本间的距离变化间接反映数据分布的变化。未来工作中将把该漂移检测方法集成到在线学习器中,以解决时间序列有关的实际问题。

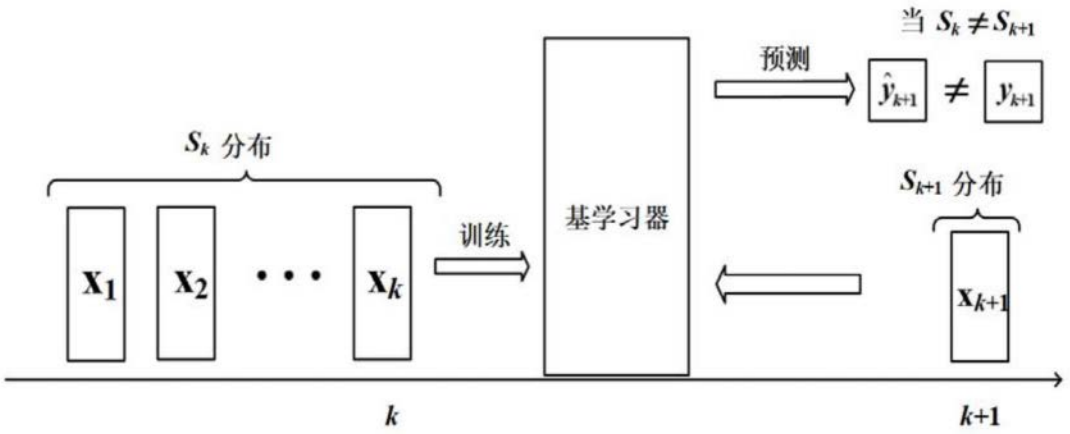

当 $S_k \neq S_{k+1}$
预测
$\hat{y}_{k+1}$
$\neq$
y_{k+1}
S_k 分布
S_{k+1} 分布
训练
基学习器
x_1
x_2
x_k
x_{k+1}
k
$k+1$

图1

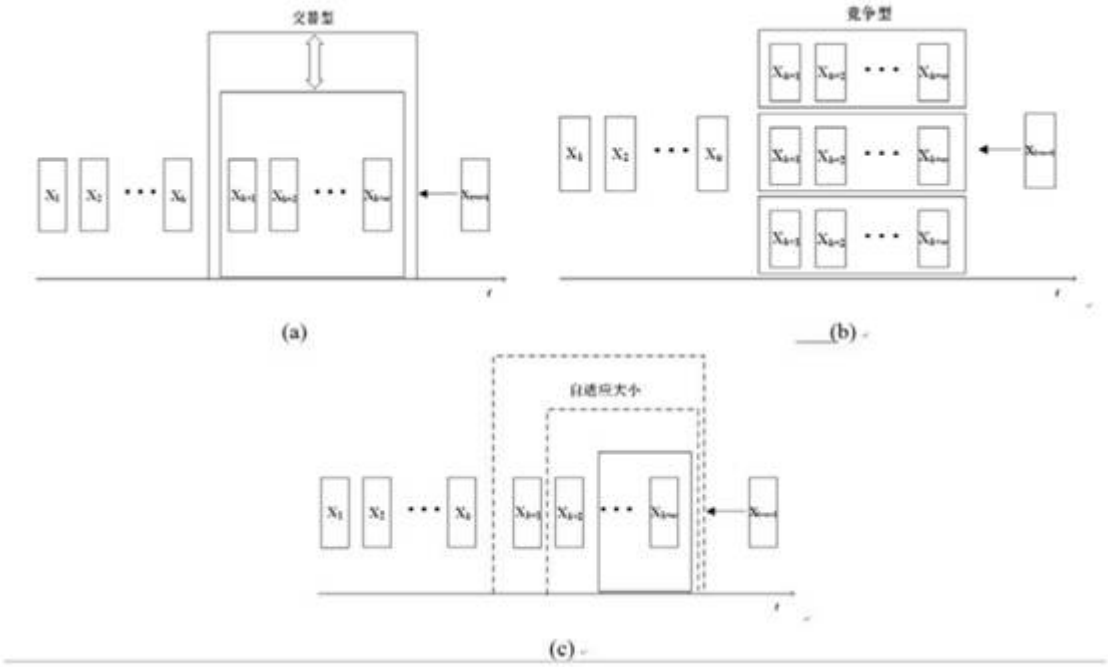

交替型
竞争型
自适应大小
(a)
(b)
(c)

图2

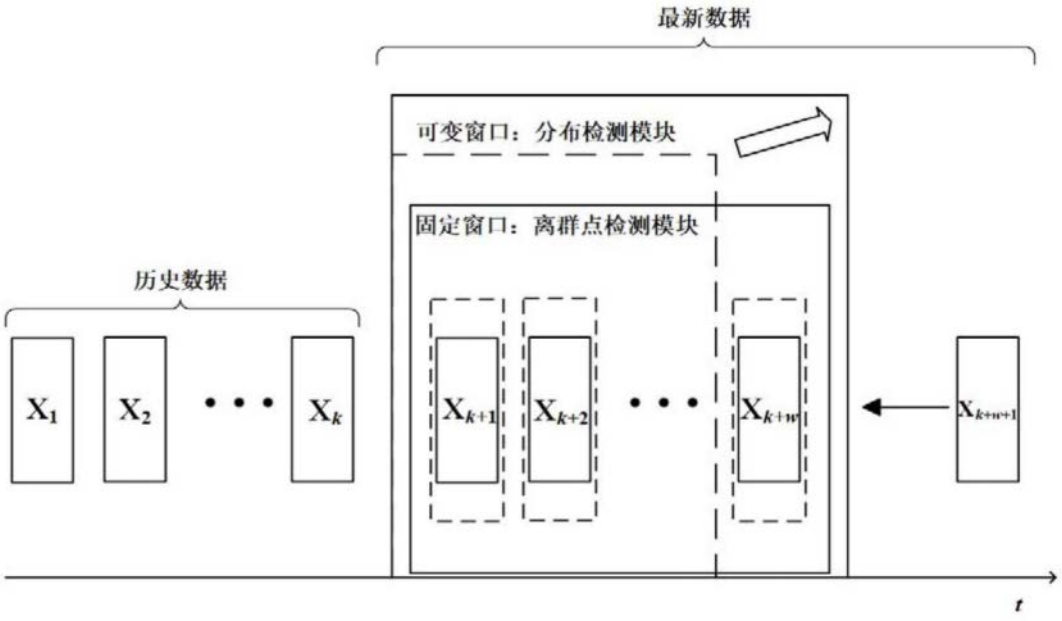


图3

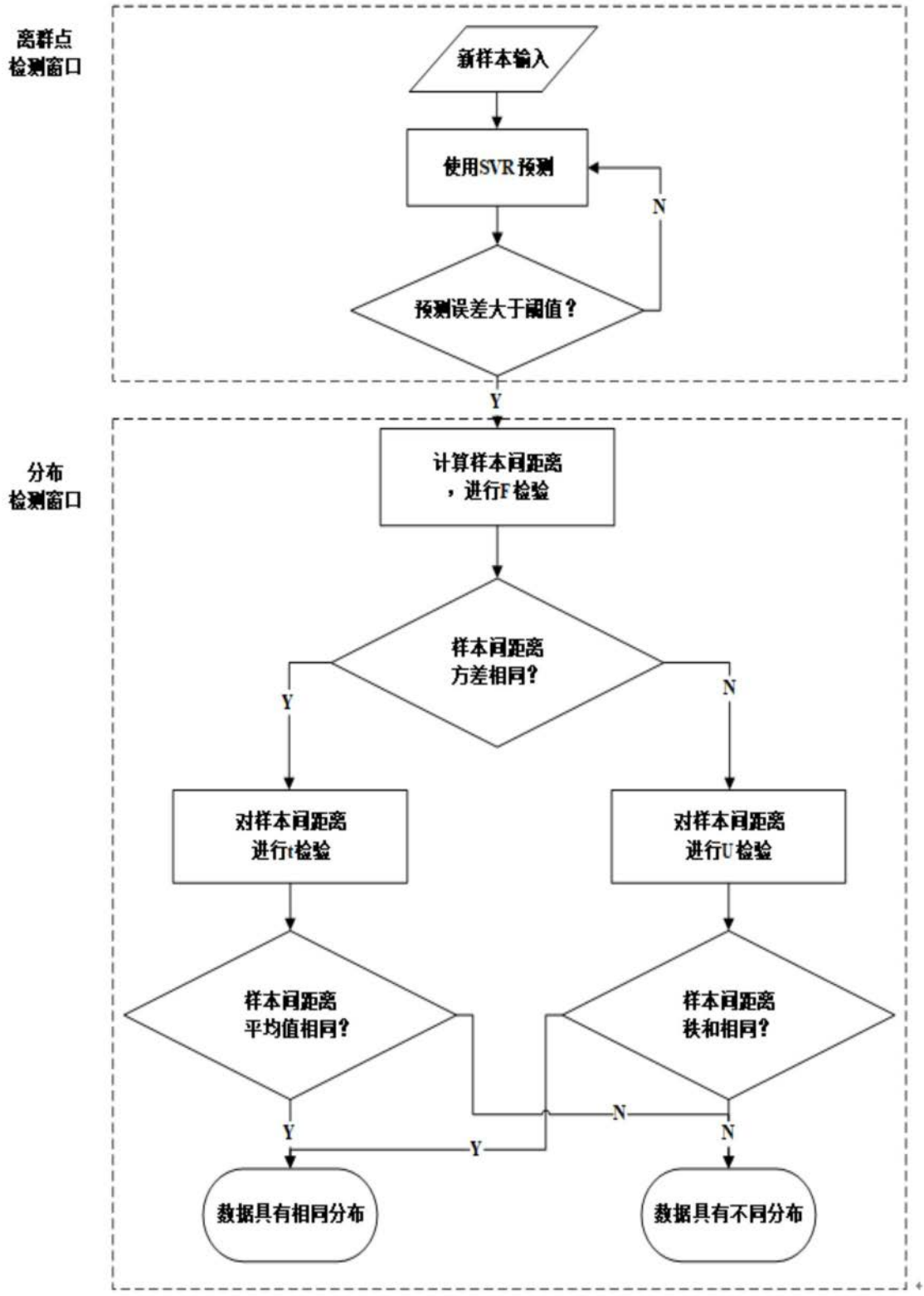

离群点
检测窗口
新样本输入
使用SVR预测
N
预测误差大于阈值？
Y
分布
检测窗口
计算样本间距离
，进行F检验
样本间距离
方差相同？
Y
N
对样本间距离
进行t检验
对样本间距离
进行U检验
样本间距离
平均值相同？
样本间距离
秩和相同？
N
Y
Y
N
数据具有相同分布
数据具有不同分布

图4

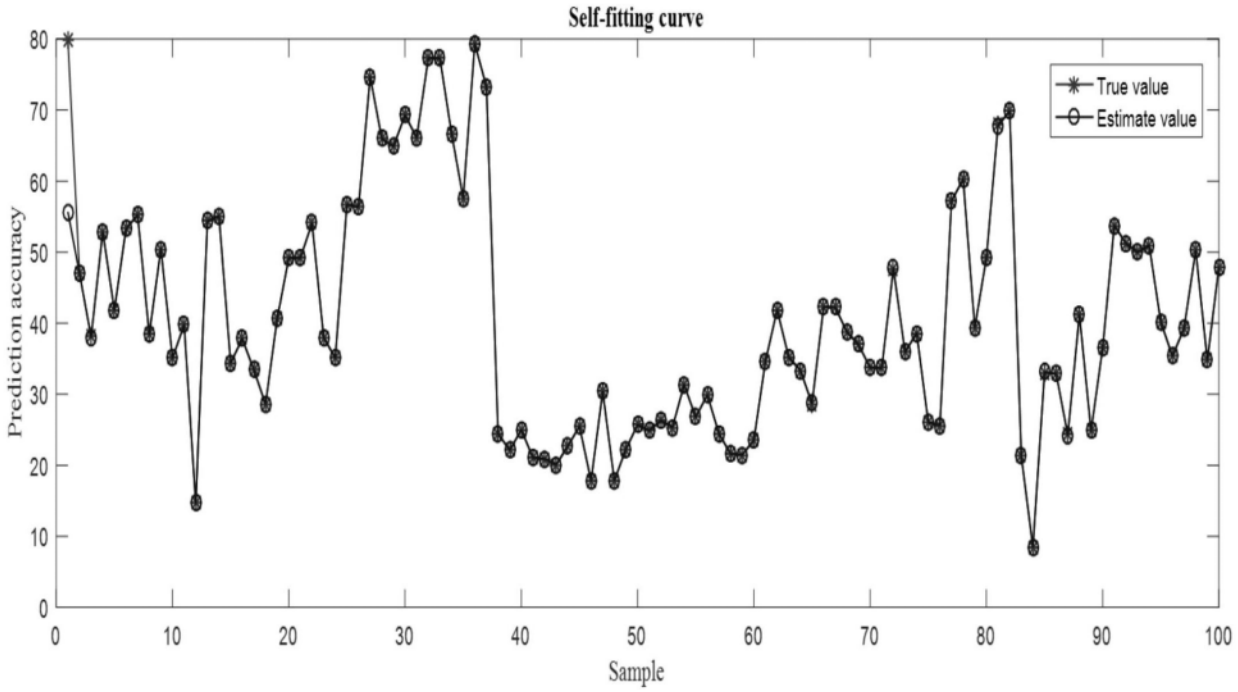
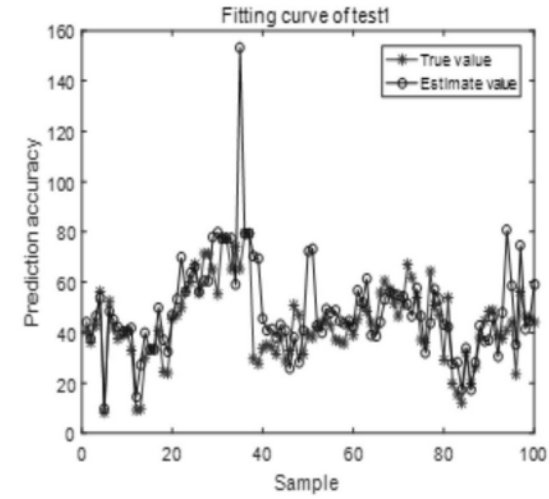
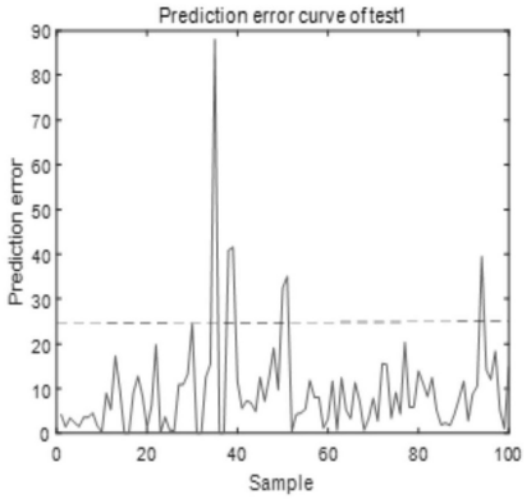

Self-fitting curve
True value
Estimate value
Prediction accuracy
Sample
0
10
20
30
40
50
60
70
80
90
100

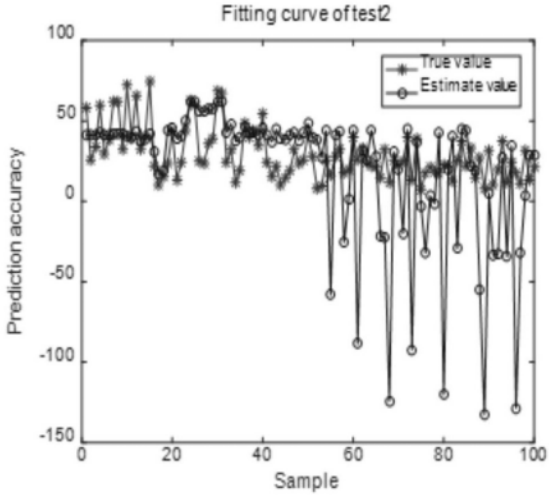
图5


Fitting curve of test1
True value
Estimate value
Prediction accuracy
Sample

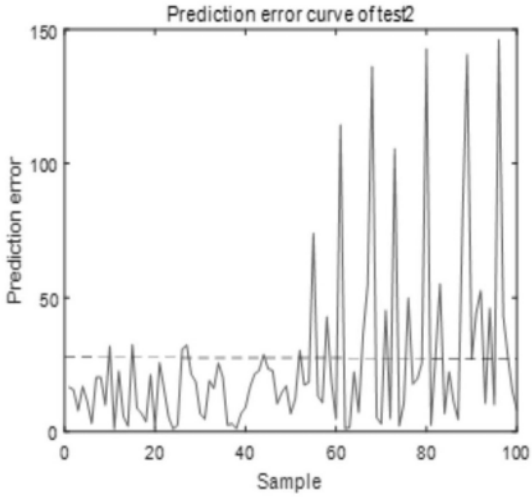
(a)


Prediction error curve of test1
Prediction error
Sample

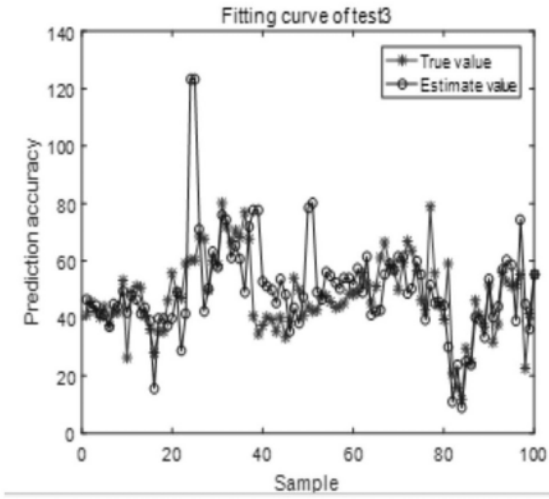
(b)


Fitting curve of test2
True value
Estimate value
Prediction accuracy
Sample

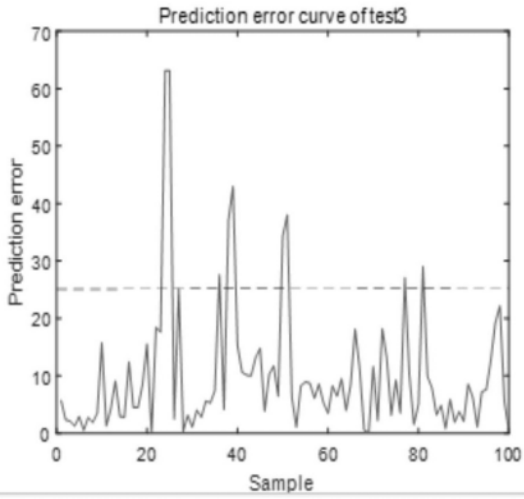
(c)


Prediction error curve of test2
Prediction error
Sample

(d)


Fitting curve of test3
True value
Estimate value
Prediction accuracy
Sample

(e)


Prediction error curve of test3
Prediction error
Sample

(f)

图6

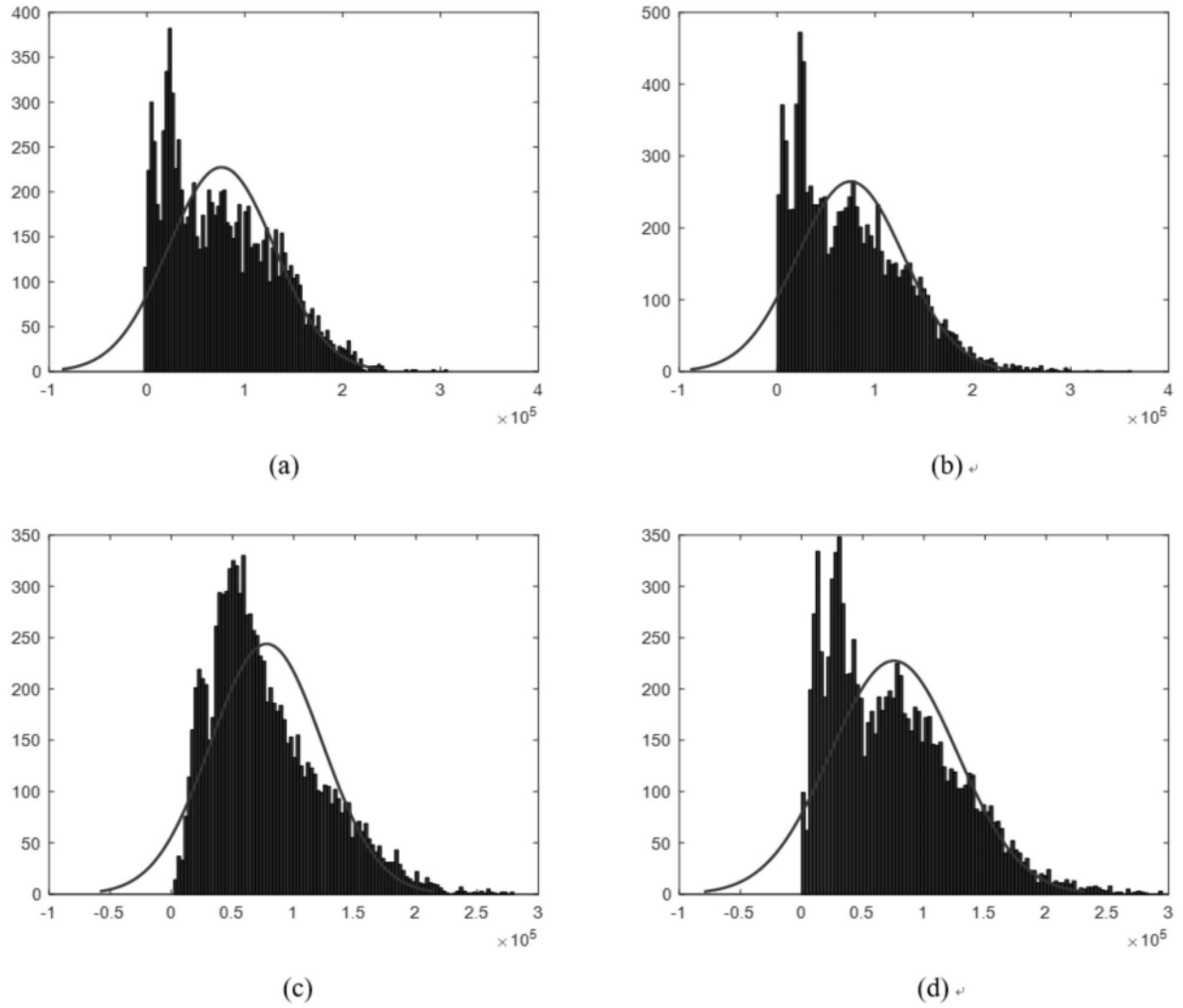


图7

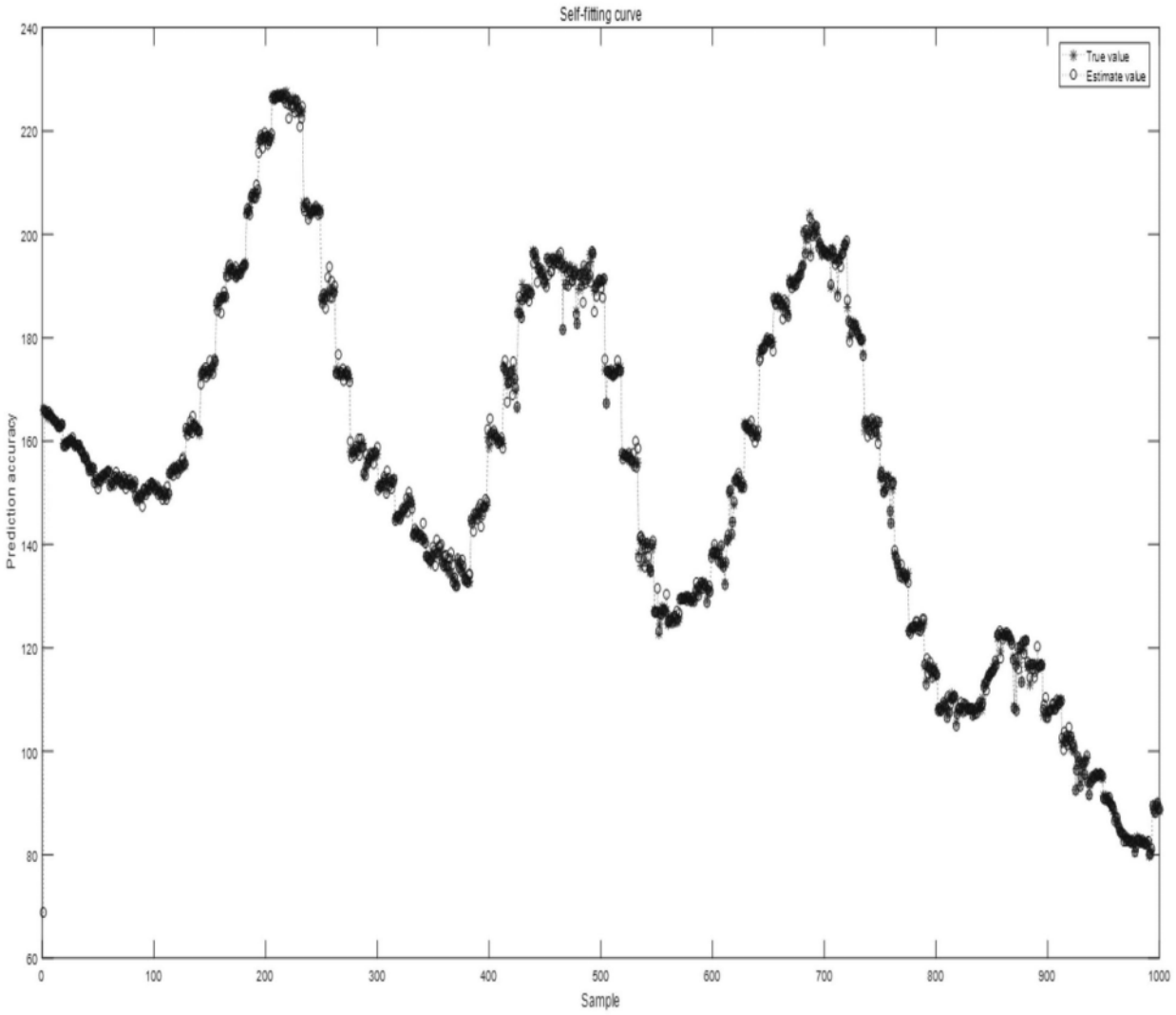

Self-fitting curve
True value
Estimate value
Prediction accuracy
Sample
240
220
200
180
160
140
120
100
80
60
0
100
200
300
400
500
600
700
800
900
1000

图8

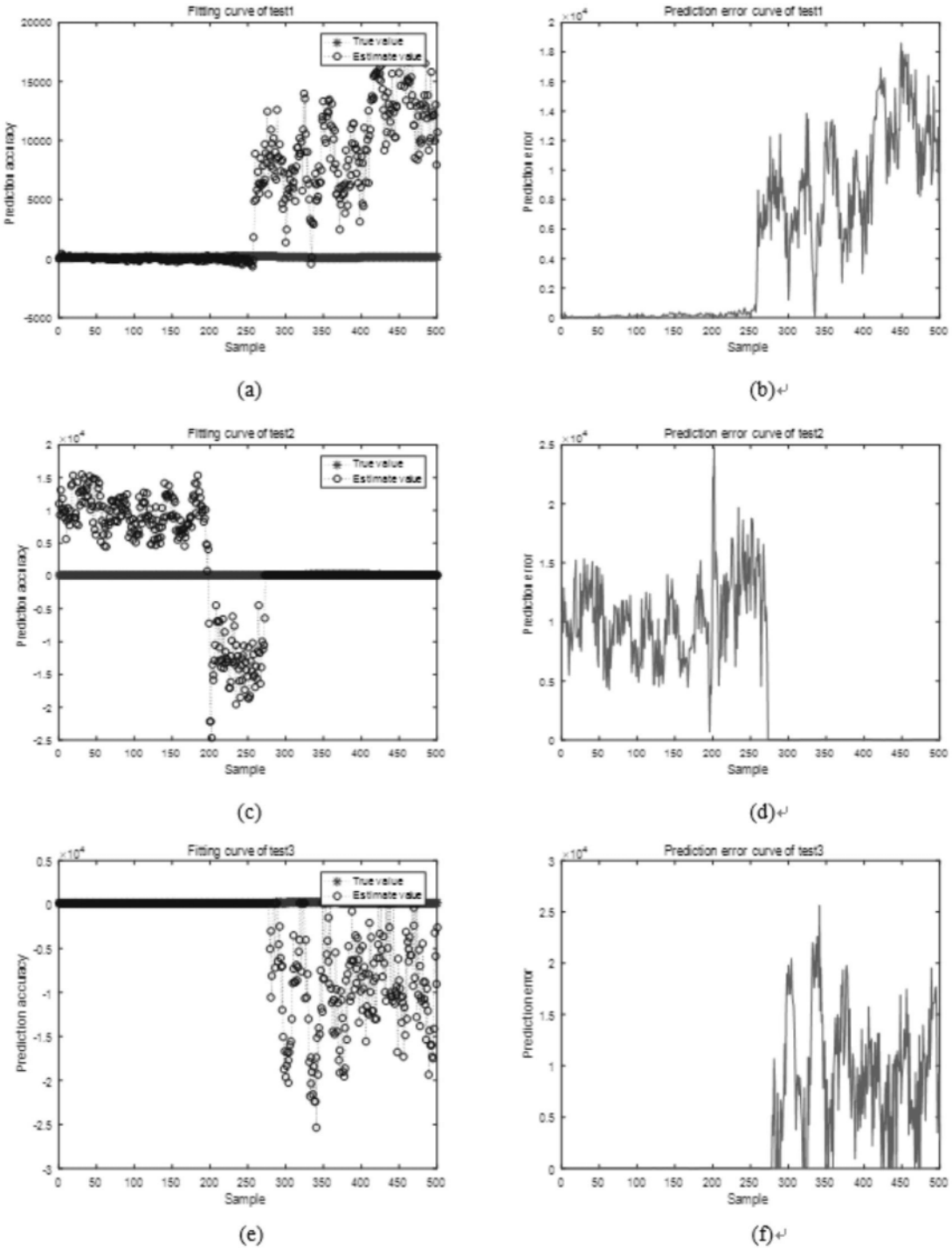

Fitting curve of test1
True value
Estimate value
Prediction accuracy
Sample
(a)
Prediction error curve of test1
Prediction error
Sample
(b)
Fitting curve of test2
True value
Estimate value
Prediction accuracy
Sample
(c)
Prediction error curve of test2
Prediction error
Sample
(d)
Fitting curve of test3
True value
Estimate value
Prediction accuracy
Sample
(e)
Prediction error curve of test3
Prediction error
Sample
(f)

图9

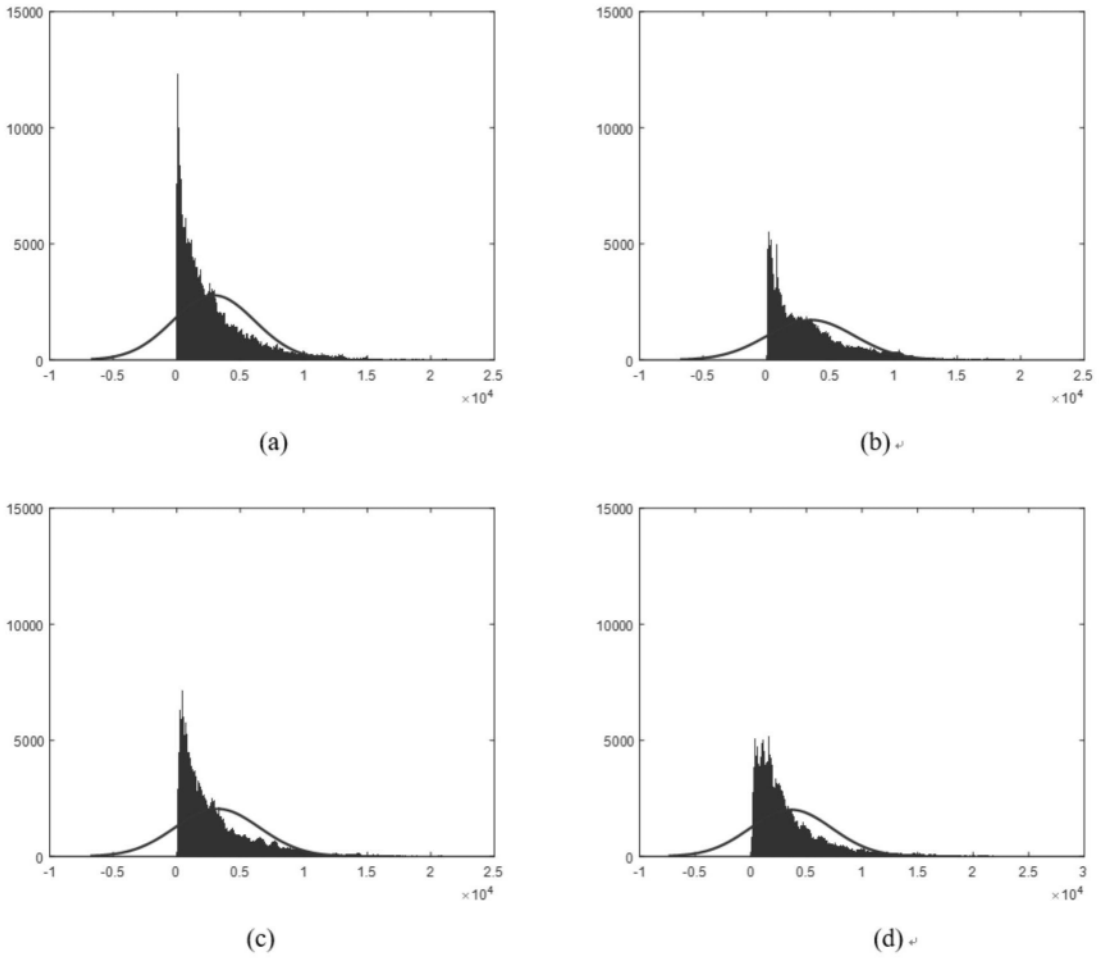


图10