基于 knnVAR 模型的地理传感数据预测

廖仁健 周丽华 肖 清 杜国王
(云南大学信息学院 昆明 650000)

摘 要 地理传感数据的预测在经济、工程、自然科学和社会科学中被广泛应用。数据中不同站点的空间相关性和同一站点的时间相关性给传统的预测方法带来了极大的挑战。文中提出了一种将数据中时间信息和空间信息有效融合,同时考虑了各传感序列独特性的 knnVAR 模型,来对地理传感数据进行预测。该模型通过计算时空距离量化数据中的时间信息和空间信息,并基于时空距离寻找 K 近邻,最后再将近邻结果应用于向量自回归模型中完成预测。knn-VAR 模型采用寻找时空近邻的方式将数据中时间维度和空间维度的相关性进行有效融合,同时使用在时空上具有高度相关性的近邻对传感序列进行预测,充分考虑了各地理序列的独特性。实验结果表明,knnVAR 模型能有效提高地理传感数据的预测精度。

关键词 地理传感数据,时空距离,K 近邻,向量自回归模型
中图法分类号 TP301 **文献标识码** A

Prediction of Geosensor Data Based on knnVAR Model

LIAO Ren-jian ZHOU Li-hua XIAO Qing DU Guo-wang
(School of Information Science & Engineering, Yunnan University, Kunming 650000, China)

Abstract The prediction of geosensor data is widely used in economy, engineering, natural science and social sciences. The spatial correlation of different sites and the time correlation of the same site in the data pose great challenges to traditional forecasting models. In this paper, a knnVAR model which computes the relevance of the space-time information effectively and considers the uniqueness of each sensing sequence at the same time was proposed to predict the geosensor data. This model quantifies the time information and spatial information of the data by calculating the space-time distance, and then searches for the K nearest neighbor based on space-time distance. Finally, the nearest neighbor sequences were applied to the vector autoregressive model. By searching for space-time nearest neighbors, knnVAR model computes the relevance of the time dimension and space dimension effectively. At the same time, knnVAR model uses the space-time nearest neighbor sequences which are highly correlated to predict the sensing sequence. The experimental results show that the knnVAR model can improve the prediction accuracy of geosensor data effectively.

Keywords Geosensor data, Space-time distance, K nearest neighbor, Vector autoregressive model

1 引言

随着地理传感网络的发展,人们获取信息的能力得到了不断提升,所获取的地理传感数据也越来越多(如气象数据、环境污染数据、交通流数据、核泄漏事件中核辐射强度数据等)。地理传感数据是一种普遍存在的时序数据,包含了时间信息和空间信息。对地理传感数据进行预测可以为决策者提供许多有用信息(如对降雨量数据的预测可以使人们提前做好防洪或抗旱的准备),有很高的实用价值。

在过去的 20 年中,学者们提出了许多对时间序列数据进行预测的算法[1-3],并取得了较好的效果,但这些模型都只考虑了数据的时间信息,并不能直接应用于具有时空特征的地理时间序列数据的预测中。对时间和空间信息同时处理要比单纯地考虑时间信息或空间信息更为复杂。地理时间序列数据中不同站点的空间相关性和同一站点数据的时间相关性给传统的预测方法带来了极大的挑战。

为了在预测中同时利用时间信息和空间信息,基于聚类的自回归积分滑动平均模型[4](Cluster based Autoregressive Integrated Moving Average Model, CArima)根据时序数据中的时间信息和空间信息计算了数据间的时间距离与空间距离,并将时间距离和空间距离综合为时空距离,采用时空聚类的方式来处理时间与空间维度之间的关系。基于时空聚类的向量自回归模型[5](spatio-temporal Cluster-based Vector Auto Regressive model, cVAR)在 CArima 模型的基础上应用多时间序列预测替代了单时间序列的预测,提高了预测的精度。然而上述模型只是将聚类结果通过一定方式统一应用到同一

本文受国家自然科学基金项目(61262069,61472346,61662086,61762090),云南省自然科学基金项目(2016FA026,2015FB114),云南省创新团队,云南省高校科技创新团队(IRTSTYN),云南大学创新团队发展计划(XT412011),云南大学谱传感和边疆安全重点实验室(C6165903)资助。
廖仁健(1991-),男,硕士生,主要研究方向为数据挖掘;**周丽华**(1968-),女,博士,教授,主要研究方向为数据挖掘、社会网络分析,E-mail:lhzhou@ynu.edu.cn(通信作者);**肖 清**(1975-),女,硕士,讲师,主要研究方向为数据挖掘;**杜国王**(1994-),男,硕士生,主要研究方向为数据挖掘。

个簇中的所有序列之中(CArima 模型根据时空聚类结果为每一个簇确定统一的自回归积分滑动平均模型[6](AutoRegressive Integrated Moving Average,ARIMA)参数;cVAR 模型则根据时空聚类结果为每一个时间序列扩展 5 个序列),忽略了各个时间序列的独特性。图 1 为一时空聚类结果图,图中的点被聚为 A,B,C,D 4 个簇,点的位置为其空间位置。图中 a_1,a_2 两点都属于簇 A,但它们之间相差较大,a_1 更靠近 B 簇,而 a_2 点邻近 D 簇。将 A 簇的特征统一应用到 a_1,a_2 两点进行预测时并不能体现 a_1 和 a_2 两点的独特性,这种现象可能会影响 a_1 和 a_2 两点的预测效果。

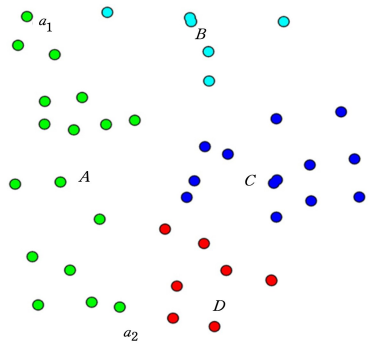


图 1 聚类分布图

为此,本文提出了 knnVAR 模型。knnVAR 模型首先根据地理时间序列数据中的空间信息和时间信息分别得到时间序列间的空间距离和时间距离,通过综合空间距离和时间距离得到序列间的时空距离。而后寻找每一个时间序列在时空上的 K 个近邻,将原序列与它的 K 个近邻序列输入向量自回归模型[7](VAR)中进行预测。knnVAR 模型中采用 K 近邻的方式替代已有模型的聚类方式,每个序列的 K 近邻是有差异的,在寻找 K 近邻时也能更好地体现每个时间序列的独特性。由于寻找 K 近邻时近邻的个数不确定,因此本文采取了抽样预测的方法来选取,利用多变量预测模型进行预测。

本文第 2 节简述与本文相关的研究工作;第 3 节介绍 VAR 模型;第 4 节描述了 knnVAR 模型,并给出具体的预测算法;第 5 节为算法时间复杂度分析;第 6 节为实验及其结果;最后总结全文。

2 相关工作

过去,地理时间序列往往被分开处理,忽略了空间的相关信息。近年来,学者们开始重视空间相关信息,并考虑如何将它应用到预测模型中。首先将空间信息应用到预测模型中的是 STARIMA 模型[8],它把一个地理时间序列看作过去观察值的线性组合,且该组合受到周边地理时间序列的影响。STARIMA 模型充分体现了距离越近的站点之间的相互影响越大这一观点。Pokrajac[9] 则应用空间自回归模型进行时空预测,将空间信息融入预测。Saengseedam 等[10] 则提出了线性混合模型进行预测,将空间信息放入条件贝叶斯框架中。以上模型在处理时都认为数据的空间相关性不变,然而数据的空间相关性与空间的潜在结构有关,即使在同一传感网络,在数据不同的情况下空间相关性往往也是不同的。Pravilovic 等[4] 则通过时空聚类的方法克服了该问题。

在应用时空聚类方法时,不同的模型也采用了不同的方式。Qin 等[11] 提出了距离测度的方式进行时空聚类,其中包含了时间距离的测度和空间距离的测度。他们合并了时空距离,而后用模糊 C 均值的方法进行聚类。Birant 等[12] 对基于密度的聚类方法进行了扩展,以处理时间与空间的关系。Appice 等[13] 提出了 SUMATRA,将地理时间序列按时间窗口的方式进行分割,根据时间窗口变化趋势的相似性对空间地理数据进行聚类,并提出了对增量数据的处理方法[14]。Pravilovics 等[5] 提出的方法中先将时间维度与空间维度分开,各自计算距离矩阵,然后把时间距离和空间距离相加得到时空距离,最后应用 PAM[15] 算法进行聚类,并将聚类结果应用到预测过程中。然而,通过聚类的方式对时间序列进行处理并不能很好地体现各个序列的独特性,本文应用寻找近邻的方式对此进行了改进,提出了 knnVAR 模型。

3 VAR(向量自回归)模型[5]

VAR 模型是由克里斯托弗·西姆斯(Christopher Sims)于 1980 年提出的计量经济学模型,自提出便被广泛应用于环境科学、气候、经济等多个领域,它能对多个时间序列组成的系统进行预测。设 V 是一个由多个时间序列组成的多变量系统,其中包含了 m 个时间序列,T 为时间窗口的大小,每个时间点用 $t=1,2,\cdots,T$ 表示,则有 $V=\{v_1(t),v_2(t),\cdots,v_m(t)\}$。VAR 模型将同一个样本期间的 m 个时间序列值当作它们过去值的线性函数。即:

$$\begin{bmatrix} v_1(T) \\ v_2(T)) \\ \vdots \\ v_m(T) \end{bmatrix} = c + A_1 \begin{bmatrix} v_1(T-1) \\ v_2(T-1)) \\ \vdots \\ v_m(T-1) \end{bmatrix} + A_2 \begin{bmatrix} v_1(T-2) \\ v_2(T-2)) \\ \vdots \\ v_m(T-2) \end{bmatrix} + \cdots + A_p \begin{bmatrix} v_1(T-p) \\ v_2(T-p)) \\ \vdots \\ v_m(T-p) \end{bmatrix} + e \tag{1}$$

其中,c 为 $m\times1$ 的常数向量,A_p 是 $m\times m$ 的参数矩阵,e 是 $m\times1$ 的误差向量。p 为滞后阶数,代表时间序列的当前值受到其 p 个过去值的影响。模型应用最小二乘法对参数 c,e,A_p 进行估计,将得到的参数用于预测未知数据,具体预测方法见 4.2 节。

VAR 模型自提出后,学者们对其进行了不断的改进,相继出现了能明确给出分量序列间的同步线性相关性的 SVAR 模型和与贝叶斯模型相结合的 BVAR 模型。然而这些模型都不能直接应用到地理传感时序数据的预测中。本文在 VAR 模型的基础上提出了能充分考虑各数据序列独特性且将数据中时间维度信息和空间维度信息融合处理的 knnVAR 模型。

4 knnVAR 模型

已有模型通常以时空聚类方式来表现数据间的时空关系。在对时间序列进行预测时将同一簇中的序列进行统一处理,这样的处理方式忽略了各个序列的独特性,在部分情况下难以获得较好的预测结果。因此,本文提出了 knnVAR 模型,通过寻找原序列的 K 近邻序列的方式将每个序列的独特性应用到预测过程中。knnVAR 模型的流程图如图 2 所示。

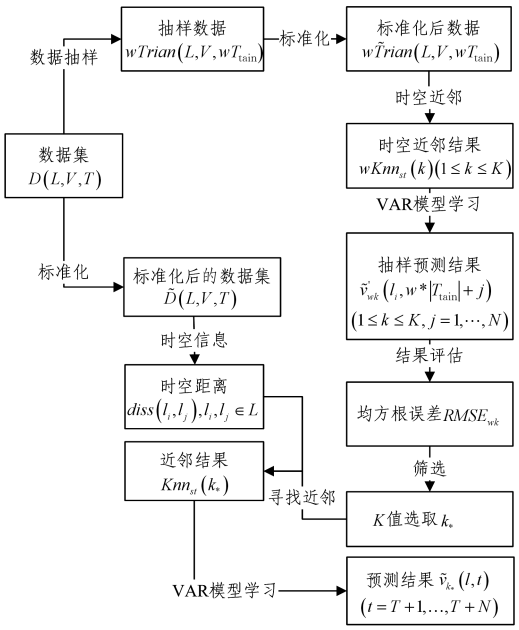


图 2　knnVAR 模型的流程图

$D(L,V,T)$为时间序列数据集,L 表示数据集中空间位置的集合,V 为观察变量,T 为时间窗口的大小,每个时间点用 $t=1,2,\cdots,T$ 表示。$v(l_i,t)$表示给定位置 $l_i\in L$ 所测得的变量 V 的序列值,空间位置 l_i 的空间坐标为(x_i,y_i)。$wTrian(L,V,T_{tain})$为选取 k 值时所抽取的训练集。$\widetilde{D}(L,V,T)$为标准化后的数据集,k_* 抽样测试选取的 k 值。$Knn_{st}(k_*)$表示指定近邻的数量为 k_* 时对数据集中每一个时间序列寻找近邻的结果,$\tilde{v}_{k_*}(l,t)$为最终预测结果,N 为预测长度。

knnVAR 模型首先根据数据集中数据的时间信息与空间信息分别得到序列间的时间距离和空间距离,将时间距离和空间距离相加得到时空距离;通过抽样测试选取得到 k_*,然后寻找各个序列的 k_* 近邻;应用原序列和该序列的近邻进行多变量的自回归模型预测;最后在预测结果中筛选出最后的预测结果。knnVAR 模型通过计算时空距离的方式将数据中的时间信息和空间信息融合,捕捉蕴含在时间序列中的空间信息和时间信息。在寻找 K 近邻的过程中寻找出各时间序列的独特性,用于提高多时间序列预测模型的预测精度。

4.1　时空 K 近邻

在寻找每个时间序列的 k 个时空近邻时,本文沿用了 cVAR 模型中的距离度量方式。首先将数据集 $D(L,V,T)$规范化为$\widetilde{D}(L,V,T)$,$\tilde{v}(l_i,t)$和$(\tilde{x}_i,\tilde{y}_i)$分别表示 $v(l_i,t)$和(x_i,y_i)的规范化值,规范化方式如式(2)和式(3)所示:

$$\tilde{v}(l_i,t)=\frac{v(l_i,t)-\langle v(l,t)\rangle_t}{\max(|v(l,t)-\langle v(l,t)\rangle_t|)} \tag{2}$$

$$\tilde{x}_i=\frac{x_i-\langle x;L\rangle}{\max(|x-\langle x;L\rangle|)},\tilde{y}_i=\frac{y_i-\langle y;L\rangle}{\max(|y-\langle y;L\rangle|)} \tag{3}$$

其中,$\langle v(l,t)\rangle_t$ 表示 K 中所有时间序列的时间均值,$\langle x;L\rangle$和$\langle y;L\rangle$分别表示 L 中所有样本横坐标和纵坐标的平均值。

在计算时空距离时,空间位置 l_i 和 l_j 所得序列之间的时空距离 $diss(l_i,l_j)$包含了序列间的时间距离 $Tdiss(l_i,l_j)$和空间距离 $Sdiss(l_i,l_j)$。$diss(l_i,l_j)$,$Tdiss(l_i,l_j)$和 $Sdiss(l_i,l_j)$的定义如式(4)－式(6)所示:

$$diss(l_i,l_j)=Tdiss(l_i,l_j)+Sdiss(l_i,l_j) \tag{4}$$

$$Tdiss(l_i,l_j)=\sum_{\mu=0}^{T-1}\alpha(1-\alpha)^{\mu}[\tilde{v}(l_i,T-\mu)-\tilde{v}(l_j,T-\mu)]^2+(1-\alpha)^{\mathrm{T}}[\tilde{v}(l_i,1)-\tilde{v}(l_j,1)]^2 \tag{5}$$

$$Sdiss(l_i,l_j)=(\tilde{x}_i-\tilde{x}_j)^2+(\tilde{y}_i-\tilde{y}_j)^2 \tag{6}$$

式(5)应用了简单指数平滑模型(Simple Exponential Smoothing,SES),$0\leqslant\alpha\leqslant1$ 为平滑参数。可以看出,α 越接近 1,就赋予最近数据更大的权重;如果 $\alpha=1$,公式蜕变为 $Tdiss(l_i,l_j)=[\tilde{v}(l_i,T)-\tilde{v}(l_j,T)]^2$,只与时刻数据 T 有关。α 的选取将在实验部分进行讨论。

根据式(4)计算得到时空距离 $diss(l_i,l_j)(l_i,l_j\in L)$后接着对每一个时间序列$\tilde{v}(l_i,t)$寻找 k 个近邻序列($\tilde{v}_{st1}(l_i,t)$,$\tilde{v}_{st2}(l_i,t),\cdots,\tilde{v}_{stk}(l_i,t)$),最终得到 $Knn_{st}(k)(1\leqslant k\leqslant K)$。如 $Knn_{st}(2)$表示每个序列的 2 近邻序列所得到的集合{($\tilde{v}_{st1}(l_1,t)$,$\tilde{v}_{st2}(l_1,t)$),($\tilde{v}_{st1}(l_2,t)$,$\tilde{v}_{st2}(l_2,t)$),$\cdots$,($\tilde{v}_{st1}(l_{|L|},t)$,$\tilde{v}_{st2}(l_{|L|},t)$)},其中 $l_1,l_2,\cdots,l_{|L|}\in L$。

4.2　多时间序列的预测

由于多变量时间序列预测模型的预测效果好于单变量时间序列预测模型的预测效果[16],因此 knnVAR 模型也应用了多变量时间序列预测模型进行预测。但是与 cVAR 模型中把每一个时间序列和其扩展序列作为一个多变量系统不同,knnVAR 模型将每个序列的原序列和它的 K 近邻序列作为一个多变量系统,从而在预测过程中充分体现了各个序列的独特性。

在得到 K 近邻结果 $Knn_{st}(k)$后,knnVAR 模型对每个序列在不同的 k 值情况下都建立如式(7)所示的 VAR 模型[7]:

$$\begin{bmatrix}\tilde{v}(l_i,T)\\ \tilde{v}_{st1}(l_i,T)\\ \vdots\\ \tilde{v}_{stk}(l_i,T)\end{bmatrix}=c+\sum_{p=1}^{P}A_p\begin{bmatrix}\tilde{v}(l_i,T-p)\\ \tilde{v}_{st1}(l_i,T-p)\\ \vdots\\ \tilde{v}_{stk}(l_i,T-p)\end{bmatrix}+e \tag{7}$$

其中,p 为滞后阶数。

应用最小二乘法对参数 $c,e,A_p(p=1,\cdots,P)$进行估计后,应用式(8)对$\tilde{v}_k(l_i,T+j)$进行预测,其中,$j=1,\cdots,N$,k 为所取近邻的个数,$1\leqslant k\leqslant K$。

$$\tilde{v}_k(l_i,T+j)=c(1)+\sum_{p=1}^{P}A_p(1,:)\begin{bmatrix}\tilde{v}(l_i,T+j-p)\\ \tilde{v}_{st1}(l_i,T+j-p)\\ \vdots\\ \tilde{v}_{stk}(l_i,T+j-p)\end{bmatrix}+e(1)(j=1,\cdots,N) \tag{8}$$

从而得到预测结果$\tilde{v}_k(l_i,T+j)$。

算法 1 为多时间序列预测算法的伪代码。

算法 1　多时间序列预测算法

输入:标准化后的地理传感时序数据集$\widetilde{\mathrm{D}}$(L,V,T),预测步数 N

输出:预测结果$\tilde{\mathrm{v}}_{\mathrm{k}}(\mathrm{l_i},\mathrm{T}+\mathrm{j})$,j=1,2,…,N,k=1,2,…,K

1. 按式(5)和式(6)分别计算序列间的时间距离 $\mathrm{Tdiss}(\mathrm{l_i},\mathrm{l_j})$和空间距离 $\mathrm{Sdiss}(\mathrm{l_i},\mathrm{l_j})$,$\mathrm{l_i},\mathrm{l_j}\in\mathrm{L}$;
2. 按式(4)计算序列间的时空距离 $\mathrm{diss}(\mathrm{l_i},\mathrm{l_j})$;
3. For k=1:K
4. 根据时空距离 $\mathrm{diss}(\mathrm{l_i},\mathrm{l_j})$得到每个序列的时空 k 近邻 $\mathrm{Knn_{st}(k)}$
5. 　　For each $\mathrm{l_i}\in\mathrm{L}$
6. 　　将原序列与其近邻结合得到($\tilde{\mathrm{v}}(\mathrm{l_i},\mathrm{t})$,$\tilde{\mathrm{v}}_{\mathrm{st1}}(\mathrm{l_i},\mathrm{t})$,$\tilde{\mathrm{v}}_{\mathrm{st2}}(\mathrm{l_i},\mathrm{t})$,…,$\tilde{\mathrm{v}}_{\mathrm{stk}}(\mathrm{l_i},\mathrm{t})$)
7. 将所得序列按式(7)学习 VAR 模型参数
8. 按式(8)得到预测结果$\tilde{\mathrm{v}}_{\mathrm{k}}(\mathrm{l_i},\mathrm{T}+\mathrm{j})$,j=1,2,…,N。

9. Endfor

10. Endfor

11. 输出预测结果 $\tilde{v}_k(l_i,T+j)$,j=1,2,…,N,k=1,2,…,K。

4.3 k 值的确定及预测

在寻找近邻序列时,近邻的个数是不确定的,即 k 的值不确定。在实际应用中即使对不同 k 值情况都进行预测,也无法判定预测结果的好坏,从而无法选择。由此,本文采用了抽样测试的方式确定 k 的值。在实验中每个数据集都被分为了训练集 $Trian(L,V,T_{tain})$和测试集 $Test(L,V,T_{test})$,$|T_{test}|=N$,应用模型对训练集进行处理得到预测结果,最后与测试集对比评估预测精度。抽样测试则是将训练集按一定比例 W 进行抽样得到抽样训练集 $wTrian(L,V,wT_{tain})$。应当注意的是,抽样时需要按时间顺序进行抽取,且相对应的抽样测试集 $wTest(L,V,wT_{test})$为抽样训练集后的数据,其长度与原测试集相同,如抽样中对 l_i 位置的时间序列抽取了 $\tilde{v}(l_i,t)$,$t=1:(w*T_{tain})$,则相对应的测试集中 l_i 位置的测试数据为 $\tilde{v}(l_i,t)$,$t=(w*T_{tain}):(w*T_{tain})+N$。在取得抽样后应用 knnVAR 模型进行预测,在不同的 k 值和抽样比例 W 下得到相应的预测结果 $v'_{wk}(l_i,w*|T_{tain}|+j)$,$1\leqslant k\leqslant K$,$j=1,\cdots,N$。利用已知的数据 $\tilde{v}(l_i,w*|T_{tain}|+j)$与预测结果 $\tilde{v}'_{wk}(l_i,w*|T_{tain}|+j)$进行对比,采用式(9)中的均方根误差 RMSE(值越小,预测精度越高)评估预测的结果。

$$RMSE_{wk}=\sqrt{\text{mean}(\tilde{v}'_{wk}(l_i,w|T_{tain}|+j)-\tilde{v}(l_i,w|T_{tain}|+j))^2} \quad (9)$$

设定不同的 k 值和 W,将得到不同的 $RMSE_{wk}$,根据式(10)找到其中的最小值 $RMSE_*$,确定其对应的 k 值 k_*。

$$RMSE_*=\min(RMSE_{wk}) \quad (10)$$

最后将 k_* 作为确定的 k 值,应用 knnVAR 模型对整个数据集 $\tilde{D}(L,V,T)$的训练集进行预测,得到预测结果 $\tilde{v}_*(l,t)$,$t=T+1,\cdots,T+N$,并用测试集评估精度,W 的取值在实验中讨论。

以下为 knnVAR 预测算法的伪代码。

算法 2 knnVAR 预测算法

输入:标准化后的训练集 Trian(L,V,T_{tain}),预测步数 N,抽样比例 W

输出:预测结果 $\tilde{v}_{k_*}(l_i,T+j)$,j=1,2,…,N

1. 按式(5)和式(6)分别计算序列间的时间距离 Tdiss(l_i,l_j)和空间距离 Sdiss(l_i,l_j),l_i,$l_j\in L$。
2. 按式(4)计算序列间的时空距离 diss(l_i,l_j)。
3. 根据 W 对数据集 Trian(L,V,T_{tain})抽样,得到抽样训练集 wTrian(L,V,wT_{tain})。
4. 应用算法 1 对抽样训练集 wTrian(L,V,wT_{tain})进行预测(预测步数为 N),得到 $\tilde{v}_{wk}(l_i,T+j)$,j=1,2,…,N,k=1,2,…,K。
5. For k=1:K
6. 根据式(9)计算得到步骤 4 中预测结果的 $RMSE_{wk}$。
7. Endfor
8. 按式(10)计算得到最小均方根误差 $RMSE_*$,同时获取相应的 k_* 值。
9. 根据步骤 2 得到的时空距离 diss(l_i,l_j)计算每个序列的时空 k_* 近邻($Knn_s t(k_*)$)。
10. For each $l_i\in L$
11. 将原序列与其近邻结合,得到($\tilde{v}(l_i,t)$,$\tilde{v}_{st1}(l_i,t)$,$\tilde{v}_{st2}(l_i,t)$,…,$\tilde{v}_{stk_*}(l_i,t)$)。
12. 将所得序列按式(7)学习 VAR 模型参数。
13. 按式(8)得到预测结果 $\tilde{v}_{k_*}(l_i,T+j)$,j=1,2,…,N。
14. Endfor
15. 输出预测结果 $\tilde{v}_{k_*}(l_i,T+j)$,j=1,2,…,N。

5 算法时间复杂度分析

knnVAR 模型首先按算法 2 中的步骤 1 和步骤 2 计算了数据集中时序数据的时间距离、空间距离和时空距离,时间耗费为$(|L|^2-|L|)(T+2)/2$。其中,$(|L|^2-|L|)/2$ 为距离矩阵中要计算的距离数量,$(T+2)$为计算标准化后计算时序数据的时间距离、空间距离和时空距离的时间耗费。综上,计算时间距离和空间距离矩阵的时间复杂度为 $O(|L|^2T)$。

算法 2 中的步骤 3 和步骤 4 的主要时间耗费为步骤 4 中的算法 1 的预测。算法 1 中步骤 1 和步骤 2 与算法 2 中步骤 1 和步骤 2 的复杂度相同,为 $O(|L|^2T)$。算法 1 中步骤 3—步骤 10 的主要耗时为 K 近邻的寻找和 VAR 模型的预测,时间复杂度为 $O(|L|^2\log|L|+|L|K^4p_{\max}^4+|L|K^3p_{\max}^3T)$,其中 K 近邻寻找采取排序方式,时间复杂度为 $O(|L|^2\log|L|)$。计算包含 m 个变量、滞后阶数为 p_i 的向量自回归模型的时间复杂度为:$O(m^3p_i^3+m^2p_i^2T)$。其中,p_i 根据 FPE 标准[3]进行选择。假设 p_i 的取值范围为 $1\sim p_{\max}$,则对 p_i 进行选择的时间耗费为:$m^3\sum_{p_i=1}^{p_{\max}}p_i^3+m^2\sum_{p_i=1}^{p_{\max}}p_i^2T$。在不同数量的近邻情况下对每一个序列进行计算,时间耗费为 $|L|\sum_{m=2}^{K+1}(m^3\sum_{p_i=1}^{p_{\max}}p_i^3+m^2\sum_{p_i=1}^{p_{\max}}p_i^2T)$,所以预测部分的时间复杂度为:$O(|L|K^4p_{\max}^4+|L|K^3p_{\max}^3T)$。

算法 2 中步骤 5—步骤 14 的主要耗时为 K 近邻的寻找和 VAR 模型的预测,时间复杂度为 $O(|L|^2\log|L|+|L|K^3p_{\max}^4+|L|K^2p_{\max}^3T)$。由此,knnVAR 模型的时间复杂度为 $O(|L|^2(T+\log|L|)+|L|K^4p_{\max}^4+|L|K^3p_{\max}^3T)$。

6 实验

6.1 实验数据

本文实验也使用了 cVAR 模型实验数据集中的数据[5],这些数据来自 4 个传感网络。将每一个数据集划分为训练集与测试集(分别用 Trian 和 Test 表示)。数据集的信息如表 1 所列。

表 1 数据集信息

传感网络	测量变量	UM	$\|L\|$	*Trian*	*Test*	Δ
TCEQ	Wind Speed	mph	26	336	24	1 h
TCEQ	Air Temperature	°F	26	336	24	1 h
TCEQ	Ozone Concentration	ppb	26	336	24	1 h
MESA	NO_X Concentration	ppb	20	268	12	2 weeks
NREL	Wind speed	m/s	1326	144	48	30 min
NCDC	Air Temperature	℃	72	93	12	1 month
NCDC	Solar Energy	MJ/m^{-2}	72	93	12	1 month

其中,UM 为测量变量对的单位,$|L|$为传感器的数量,*Trian* 为训练集的长度,*Test* 为测试集的长度,Δ 为抽样时间间隔。

4 个传感网络分别是:1)TCEQ(Texas)传感网络[1],该传感网络测量了 3 个变量,分别为风速、气温和臭氧浓度,该传感网络共有 26 个传感器,本文截取了 2009 年 5 月 5 日至 5 月 19 日的数据,每次测量时间间隔为 1 小时;2)MESA 大气污染研究传感网络[2],该传感网络有 20 个传感器,测量了氮氧化合物浓度,本文截取了 1999 年 1 月 13 日至 2009 年 9 月 23 日的数据,每次测量时间间隔为 2 周;3)NREL 传感网络[3],该传感网络有 1326 个传感器,测量的变量为风速,本文截取了 2004 年 1 月 1 日至 2004 年 1 月 4 日的数据,每次测量时间间隔为 30 min;4)NCDC 传感网络[4],该传感网络测量了两个变量,分别为气温和太阳能强度,该传感网络共有 72 个传感器,本文截取了 2005 年 8 月至 2014 年 4 月的数据,每次测量时间间隔为 1 个月。

6.2　实验结果

实验对比了 knnVAR 模型的预测结果与 cVAR 模型的预测结果,且在两个算法中都对时间距离计算中的平滑参数 α 进行了讨论,计算了快速给定 $\alpha=0.5$ 和根据数据的不同情况对 α 进行估计($\alpha=est$)时的预测结果。每个数据集都被分为训练集和测试集,在数据预处理时按照先训练集后测试集的顺序处理,而后应用不同模型对训练集进行处理,得到预测结果。对比预测结果与测试集,用式(9)中的均方根误差 RMSE 评估预测精度。我们采用了抽样测试的方法确定 knnVAR 模型中的近邻值,分析了不同抽样比例 W 时的预测结果。在多时间序列的预测中,滞后阶数 p 则根据 FPE 准则进行选取[3],$p_{\max}$为 10。由于在预测时加入周期参数后预测结果要好于不加入周期参数时的预测结果[5],因此本文在实验结果中只展示了加入周期参数后的实验结果,如表 2 所列。

表 2　实验结果

传感网络	测量变量	平均 *RMSE*					
		cVAR $\alpha=0.5$	cVAR $\alpha=$est	knnVAR $\alpha=0.5,w=50\%$	knnVAR $\alpha=est,w=50\%$	knnVAR $\alpha=0.5,w=80\%$	knnVAR $\alpha=est,w=80\%$
TCEQ	Wind Speed	0.32	0.32	0.33($k=1$)	0.33($k=1$)	0.31($k=7$)	**0.30($k=4$)**
TCEQ	Air Temperature	**0.22**	**0.22**	0.22($k=1$)	0.40($k=7$)	**0.22($k=1$)**	**0.22($k=1$)**
TCEQ	Ozone Concentration	0.53	0.54	0.52($k=3$)	**0.52($k=4$)**	**0.52($k=1$)**	0.53($k=1$)
MESA	NO_X Concentration	0.27	0.27	**0.20($k=1$)**	0.27($k=3$)	**0.20($k=1$)**	**0.20($k=1$)**
NREL	Wind Speed	0.44	0.43	**0.41($k=1$)**	**0.41($k=1$)**	**0.41($k=1$)**	**0.41($k=1$)**
NCDC	Air Temperature	**0.12**	**0.12**	**0.12($k=1$)**	**0.12($k=1$)**	**0.12($k=1$)**	**0.12($k=1$)**
NCDC	Solar Energy	0.37	0.14	**0.12($k=1$)**	**0.12($k=1$)**	**0.12($k=1$)**	**0.12($k=1$)**

实验结果表明,本文模型在对 MESA 传感网络所产生的数据集、NCDC 传感网络中的太阳能强度数据集和 NREL 传感网络所产生的数据集进行预测时所得到的预测结果的均方根误差要小于 cVAR 模型,即相比 cVAR 模型,本文模型预测的精度更高,而在其余数据集中两个模型的预测结果相差并不大。

在进行样本抽取时 w 为 80%时的预测结果要好于 w 为 50%时的预测结果,这是由于在时间序列的预测中需要有足够多的样本作为测试集,样本数量太少时预测结果的质量会有所下降。实验中 α 分别取了 0.5 与 *est* 两个值,在 cVAR 模型和 knnVAR 模型中,α 值对预测效果没有太大影响。

在实验结果中,knnVAR 模型对 TCEQ 传感网络臭氧浓度数据集($\alpha=0.5$)的预测结果要好于 cVAR 模型的预测结果,这是因为 knnVAR 模型在测试时考虑了各个序列的独特性。图 3 为 TCEQ 传感网络臭氧浓度数据集的聚类结果与 S 点五近邻展示图($\alpha=0.5$),图中横坐标为标准化经度,纵坐标为标准化纬度,各点表示 TCEQ 传感网络臭氧浓度数据集中各个传感器的空间位置。点的形状表示时空聚类结果($\alpha=0.5$),不同的簇用不同的形状符号表示,用虚线连接的点为 S 点的 5 个时空近邻(计算近邻时所用距离为时空距离,时空近邻在空间上不一定邻近)。点 S 属于圆形符号簇,但它的近邻却不全属于圆形符号簇。cVAR 模型对 S 序列进行预测时只应用圆形符号簇中信息,忽略了三角符号簇中点对 S 序列的影响。knnVAR 模型则利用近邻序列对 S 序列进行预测,考虑了 S 序列的独特性,打破了聚类带来的约束。

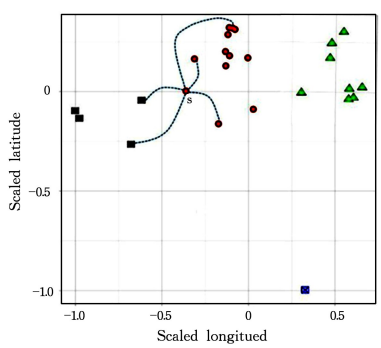


图 3　TCEQ 传感网络臭氧浓度数据集聚类结果与 S 点五近邻展示图($\alpha=0.5$)

在 k 取不同值时,将得到不同的预测结果,如图 4 是 $w=100\%$时 k 值对预测结果的影响($\alpha=est$),图中横坐标为近邻的数量 K,纵坐标为预测结果的均方根误差 *RMSE*,各条折线代表不同的数据集。图 4(a)为 TCEQ 传感网络数据,图 4(b)为 MESA,NREL 和 NCDC 传感网络数据。TCEQ 传感网络中 Ozone 数据集预测结果的 *RMSE* 值随着 k 值的增加小幅降低,Wind 数据集预测结果的 *RMSE* 值在 k 值变化时变化不大,其余数据集的预测精度都随着 k 值的增加而降

(下转第 457 页)

[1] http://www.tceq.state.tx.us/

[2] http://depts.washington.edu/mesaair/

[3] http://www.nrel.gov/

[4] https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/solar-radiation

Conf on Pervasive Computing. 2012.

[12] WANG J, PRABHALA B. Periodicity based next place prediction[C]// Nokia Mobile Data Challenge 2012 Workshop Dedicated Task. Citeseer, 2012.

[13] GAO H, TANG J, LIU H. Mobile location prediction in spatio-temporal context[C]// Nokia Mobile Data Challenge Workshop. Citeseer, 2012.

[14] BAUMANN P, KLEIMINGER W, SANTINI S. The influence of temporal and spatial features on the performance of next-place prediction algorithms[C]// Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 2013: 449-458.

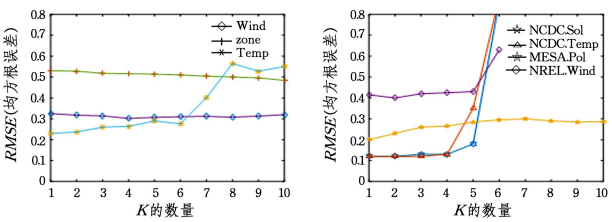
[15] TRAN L H, CATASTA M, MCDOWELL L K, et al. Next Place Prediction using Mobile Data[C]// Proceedings of the Mobile Data Challenge Workshop (MDC 2012). 2012.

[16] NOULAS A, SCELLATO S, LATHIA N, et al. Mining User Mobility Features for Next Place Prediction in Location-Based Services[C]// ICDM. Citeseer, 2012: 1038-1043.

[17] DUDA R O, HART P E, STORK D G. Pattern classification [M]. John Wiley & Sons, 1999.

（上接第 435 页）

低。即不同数据情况下 k 值与预测精度没有确定的关系，大多数情况下预测精度都随着 k 值的增加而降低。在实验时我们将近邻数量 K 的最大值设置为 10。



(a)　　(b)

图 4　K 值对预测结果的影响

结束语　本文提出 knnVAR 模型对地理时间序列进行预测，首先用寻找时空 K 近邻的方式将数据中的时间信息与空间信息融合，而后利用 VAR 模型进行多时间序列的预测。在模型中充分考虑了各个时间序列的独特性，且对近邻的数量进行了讨论，提出了抽样测试的方式来确定 k 值，以提高预测的精度。在实验中对多个实际地理传感网络中的数据进行了预测，并将预测结果与未考虑各序列独特性的 cVAR 模型的预测结果进行了对比，结果表明 knnVAR 模型比 cVAR 模型有更好的预测精度。在下一步的工作中我们将对各个序列的独特性作进一步的探讨，本文模型在寻找近邻时对整个模型的 k 值进行了确定，但每个时间序列对 k 值的选取可能会不同，需要单独讨论，这将是我们以后的工作方向。

参 考 文 献

[1] EGRIOGLU E, YOLCU U, ALADAG C H, et al. Recurrent Multiplicative Neuron Model Artificial Neural Network for Non-linear Time Series Forecasting[J]. Neural Processing Letters, 2015, 41(2): 249-258.

[2] HYNDMAN R J, KHANDAKAR Y. Automatic Time Series Forecasting: The forecast Package for R[J]. Journal of Statistical Software, 2008, 27(3): 1-22.

[3] LÜTKEPOHL H. New introduction to multiple time series analysis[M]. Springer Science & Business Media, 2005: 88-89.

[4] PRAVILOVIC S, APPICE A, MALERBA D. Integrating cluster analysis to the ARIMA model for forecasting geosensor data[C]// International Symposium on Methodologies for Intelligent Systems. Cham: Springer, 2014: 234-243.

[5] PRAVILOVIC S, BILANCIA M, APPICE A, et al. Using multiple time series analysis for geosensor data forecasting[J]. Information Sciences, 2017, 380: 31-52.

[6] BOX G E P, JENKINS G M. Time Series Analysis: Forecasting and Control[J]. Journal of Time, 2010, 31(4): 303-303.

[7] TSAY R S. Multivariate time series analysis. With R and financial applications[M]. Wiley, 2013: 1-40.

[8] KAMARIANAKIS Y, PRASTACOS P. Space-time modeling of traffic flow[J]. Computers & Geosciences, 2005, 31(2): 119-133.

[9] POKRAJAC D, OBRADOVIC Z. Improved spatial-temporal forecasting through modelling of spatial residuals in recent history[C]// Proceedings of the 2001 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2001: 1-17.

[10] SAENGSEEDAM P, KANTANANTHA N. Spatial time series forecasts based on Bayesian linear mixed models for rice yields in Thailand[C]// Proceedings of the International Multi Conference of Engineers and Computer Scientists. 2014: 1007-1012.

[11] QIN K, CHEN Y, ZHAN Y, et al. Spatial clustering considering spatio-temporal correlation[C]// International Conference on Geoinformatics. 2011: 1-4.

[12] BIRANT D, KUT. ST-DBSCAN: An algorithm for clustering spatial-temporal data[J]. Data & Knowledge Engineering, 2007, 60(1): 208-221.

[13] APPICE A, CIAMPI A, MALERBAD. Summarizing numeric spatial data streams by trend cluster discovery[J]. Data Mining and Knowledge Discovery, 2015, 29(1): 84-136.

[14] APPICE A, GUCCIONE P, MALERBA D, et al. Dealing with temporal and spatial correlations to classify outliers in geophysical data streams[J]. Information Sciences, 2014, 285(1): 162-180.

[15] REYNOLDS A P, RICHARDS G, IGLESIA B D L, et al. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms[J]. Journal of Mathematical Modelling & Algorithms, 2006, 5(4): 475-504.

[16] ZIVOT E, WANG J. Modeling Financial Time Series with S-PLUS? [M]. New York: Springer, 2006: 296.