文章编号:1004-9037(2014)01-0152-05

聚类再回归方法在机场噪声时间序列预测中的应用

夏 利 王建东 张 霞 王丽娜

(南京航空航天大学计算机科学与技术学院,南京,210016)

摘要:对于机场噪声的预测,针对绘制等值线方法预测成本高和误差较大的缺点,以及分类再回归方法中分类时缺乏可指导性标准的问题,本文提出了基于支持向量机的先聚类、再回归的时间序列的预测方法。对机场噪声时间序列的先聚类再回归方法,采用常用 k 均值划分算法,利用聚类特点,将样本限定在同一类的范围内,再对同类样本进行回归预测。Housing 及 Laser generated data 数据集上的实验表明,采用先聚类再回归方法得到的拟合值比直接回归方法得到的拟合值要精确。将该方法应用到北京某机场实测数据中,并与其他预测模型进行对比,准确度明显优于其他预测方法。

关键词:支持向量机;时间序列;机场噪声;聚类;回归

中图分类号:TP399 **文献标志码**:A

Application of Cluster Regression in Time Series Prediction of Airport Noise

Xia Li, *Wang Jiandong*, *Zhang Xia*, *Wang Lina*

(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China)

Abstract: For airport noise prediction, aiming at the high cost and large error of contour drawing, as well as the lack of guidance standard in regression method based on SVM classification, it presents a method of cluster regression based on support vector machine (SVM). Cluster regression in prediction of airport noise, using k-means algorithm, takes advantage of the characteristics of clustering. It firstly limits the sample within the same class, and then performs regression in the similar class. Experimental results on housing data set and Laser generated data set show that the fitted values of the cluster regression method are more accurate than the direct regression method. Applied the method to measured data of an airport in Beijing, and compared it with other prediction models, the accuracy of cluster regression is superior to that of other prediction methods.

Key words: support vector machine (SVM); time series; airport noise; cluster; regression

引 言

机场噪声的问题随着我国民航事业的发展而日渐严重,目前对机场噪声预测的方法比较流行是采用国际民航组织推荐的指标——计权等效连续感觉噪声级绘制噪声等值线方式和以飞机的噪声距离曲线为核心,用一定的数学模型将其修正至与具体机场环境条件相关的噪声传播模型,存在预测成本高和误差较大等缺点。

文献[1]提出了一种基于 SVM 的先分类再回归计算方法,经验证,与直接回归相比,预测效果有很大改进。不过对于分类界限不明确的情况,采用先分类的方法明显有一定的局限性。针对这种情况,本文提出一种先聚类再回归的支持向量回归方法,利用聚类使结果簇内的相似度高,簇间的相似度低的优点,将测试样本限定在同一类的范围内,再对同类样本进行支持向量回归,使

基金项目:国家自然科学基金重点(61139002)资助项目。
收稿日期:2012-05-13;**修订日期**:2012-09-28

用机场已有的历史噪声监测数据进行机场噪声预测的计算方法。

1　基本理论

1.1　支持向量回归机

给定具有 l 个输入/输出的训练集 $T=\{(x_1,y_1),\cdots,(x_l,y_l)\}$，构造回归函数

$$f(x)=\boldsymbol{w}^{\mathrm{T}}\cdot\varphi(x)+b \tag{1}$$

式中：$\boldsymbol{w}$ 为权重，b 为偏置项。Vapnik 引入了一个 ε 不敏感损失函数作为损失函数，将 SVM 推广到回归问题，提出支持向量回归，线性 ε 不敏感损失函数定义如下

$$c(x,y,f(x))=|\ y-f(x)\ |_{\varepsilon} \tag{2}$$

式中 $|y-f(x)|_{\varepsilon}=\max\{0,|y-f(x)|-\varepsilon\}$，这里，$\varepsilon$ 是一个事先取定的一个正数。将式(2)限定在式(1)中估计回归函数，基于结构风险最小化原则，就得到了对回归问题的线性支持向量机算法，它要解决一个原始优化问题

$$\min\{\frac{1}{2}\|w\|^2+\frac{1}{2}\sum_{i=1}^{l}\varepsilon_i+\varepsilon_i^*\},\ i=1,2,\cdots,l$$

$$\text{s.t.}\quad ((w.x_i)+b)-y_i\leqslant\varepsilon+\xi_i^*,$$

$$y_i-((w.x_i)+b)\leqslant\varepsilon+\xi_i^*$$

$$\xi_i^*\geqslant 0 \tag{3}$$

根据 Wolfe 对偶定义，并引入核函数 $K(x,x')$ 代替内积 (x_i,x_j)，式(3)转换为如下最优问题[2,3]

$$\min\{\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(K(x_i,x_j)+\frac{1}{C}\sigma_{ij})+\varepsilon\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i)\}$$

$$\text{s.t.}\quad \sum_{i=1}^{l}(\alpha_i-a_i^*)=0$$

$$a_i,a_i^*\geqslant 0,\ i=1,2,\cdots,l \tag{4}$$

1.2　聚类分析

聚类是将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程，由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。

大体上，主要的聚类方法可以分为几类：划分方法、层次方法、基于密度的方法、基于网格的方法、基于模型的方法。将它赋给最近的簇；然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。有代表性地，平方误差准则被采用，这个准则试图使生成的结果尽可能的紧凑和独立。

2　先聚类再回归算法

2.1　先聚类再回归算法建模步骤

步骤 1　采用聚类分析中典型的划分方法：k 均值算法，将训练集 S 聚类为 k 个簇。

其中 $S=\bigcup_{i=1}^{k}S_i$，$S_i\cap S_j=\varnothing(i,j=1,2,\cdots,k,i\neq j)$

步骤 2　对于新输入预测样本，根据步骤 1 得到的聚类结果，确定样本所属类别。

(1)分别计算每一类的类中心 $C_i(i=1,2,\cdots,k)$。

(2)采用欧式距离计算样本到每一类的类中心 C_i 距离 $\mathrm{Dist}_i(i=1,2,\cdots,k)$，按到各类距离最小进行类别判定。

步骤 3　根据步骤 2 的分类结果，对属于同一类别 l 的样本进行回归预测。

(1)归一化训练样本和预测样本，减小样本之间数量级的差异。

(2)对属于同一类别 l 的训练样本，选取 S_l 为训练集，利用 ε-svr 模型进行训练，得出训练模型。

(3)对预测样本集中属于类别 l 的样本 (x_i,y_i)，用上一步得到的支持向量回归模型，对 y_i 值进行回归预测。

2.2　在常用数据集上的实验

为验证本方法的可行性和有效性，本文使用 UCI 数据库中的 Housing 数据集和 The Santa Fe Time Series Competition Data 中 Laser generated data 数据集进行实验。

由于 Vapnik 等人已经证明，不同核函数对 SVM 性能影响不大，核函数的参数和惩罚因子 C 是影响 SVM 性能对的关键因素，因此，以下 ε-svr 直接回归和聚类再回归所用核函数均为径向基函数(Radial basis function，RBF)。

关于 SVM 参数的优化选取，国际上并没有公认统一的最好的方法。目前常用的方法就是让 c 和 g 在一定的范围内取值，对于取定的 c 和 g，把训练集作为原始数据集利用 K-CV 方法得到在此组 c 和 g 下训练集验证均均方根误差(Mean Square Error，MSE)，最终取使得平均均方误差最小的那组 c 和 g 做为最佳的参数，但有一个问题就是可能会有多组的 c 和 g 对应于最小的平均均方根误差。

针对这一问题,本文采用的手段是选取能够达到最小的平均均方根误差中参数 c 最小的那组 c 和 g 做为最佳的参数,如果对应最小的 c 有多组 g,就选取搜索到的第一组 c 和 g 做为最佳的参数。这样做的理由是:过高的 c 会导致过学习状态发生,即训练集平均均方根误差很小而测试集平均均方根误差较大(泛化能力降低),所以在能够达到最小的平均均方根误差中,所有的成对的 c 和 g 中认为较小的惩罚参数 c 是更佳的选择对象。

(1) Housing 数据集

选取前 490 个样本为训练样本、10 个为测试样本,由文献[4],按训练样本的决策属性分为 5 类(1～10 为 1 类,10～15 为 2 类,16～25 为 3 类,26～50 为 4 类)在聚类时采用 k 均值方法将样本划分为 5 簇,对直接回归和聚类再回归两种方法的结果进行比较,结果如表 1 所示。两种预测方法精度比较如表 2 所示。

表 1 房屋价格

Table 1 Housing price

实际值	ε-svr 预测值	聚类再回归预测值
23.1	24.794 6	20.374 2
19.7	20.131 6	19.581 6
18.3	20.161 5	19.201 7
21.2	21.982 9	20.566 7
17.5	19.82	18.895 9
16.8	20.280 1	19.675 8
22.4	24.465 5	22.534 9
20.6	21.011 5	19.159 8
23.9	26.705	22.397 2
22	25.092 8	21.442 4

表 2 两种预测方法精度比较

Table 2 Comparison of two methods′ accuracy

度量值	ε-svr	聚类再回归
RMSE	2.156 4	1.534 2
MAPE	9.437 8	6.182 6

本文中,RMSE 为均方根误差,MAPE 为平均百分误差。

(2) Laser generated data 数据集

选取前 507 个数据,根据文献[5-6],单变量时间序列的建模方法,选择时间延时为 1,嵌入维数 7,对此时间序列做相空间重构 $X_n=\{x_n,x_{n-1},\cdots,x_{n-6}\}$,形成 7 维状态空间,以 x_{n+1} 作为 y_n,则新的时间序列为(x_i,y_i)数据对$(i=8,9,\cdots,506,507)$,选取前 490 个数据对为训练集,以最后 10 个数据对为测试集,此处聚类数目定为 2,对直接回归和聚类再回归两种方法的结果进行比较,结果如表 3 所示。两种预测方法精度比较如表 4 所示。

表 3 激光数据

Table 3 Laser data

实际值	ε-svr 预测值	聚类再回归预测值
66	57.999 9	62.418 9
237	197.203 7	210.044 9
137	118.425 0	119.151 7
25	35.650 1	28.569
9	22.769 7	12.391 8
6	20.847 4	13.999 9
5	19.634 8	5.225 2
4	17.123 5	−3.190 0
5	14.583 0	3.671 8
25	124.505 4	24.031

表 4 两种预测方法精度比较

Tab. 4 Comparison of two methods′ accuracy

度量值	ε-svr	聚类再回归
RMSE	35.910 8	10.957 5
MAPE	169.599 3	42.981 7

通过以上两个实验可以看出,通过先聚类再进行回归,可以利用聚类的优势,使测试样本的值通过具有较高相似的训练样本进行训练建的模型进行预测,从而使预测的精确度有明显提高。通过在一般的回归数据集 Housing 和时间序列数据集 Laser generated data 数据集上进行的对比试验,可以发现此方法对于回归问题相对于直接回归预测都具有优越性。

3 算法在机场噪声预测中的应用

3.1 数据选取

已有数据为北京某机场监测点 2 月至 8 月实测数据,数据为 15 个观测点每秒采集一次获得。面对大批量数据,选取 2 号和 12 号观测点,3 月至

6 月共 122 天数据。由于机场噪声具有声级高、间断性等特点，所以对一段时间内的平均值进行分析计算。

本文中，取每天 19：00～22：00 数据，每 10 min 数据求平均值，组成 18 维输入向量 $\boldsymbol{X}$，以每天 22：00～22：10 分平均值作为输出值 $\boldsymbol{Y}$。对数据划分训练集和测试集如下：选择前 115 天数据作为训练集，以最后一个星期数据作为测试集。

3.2　建模预测

(1)用 k 均值算法对训练集进行聚类

由于对机场噪声一段时间内的平均值构成较大影响的主要因素包括航班数和天气等诸多因素，考虑样本数量，不推荐簇数量过大，本文采取方法为将聚类数量从 2 递增至 5，选取使均方根误差最小的聚类数量。

(2)用 ε-svr 对各类分别进行回归预测

首先，对训练样本和测试样本的输入向量 $\boldsymbol{X}$ 进行归一化，然后，对训练集进行回归模型中惩罚因子 C 和核函数参数 σ 的寻优，再对训练集进行训练，得到支持向量回归模型，并用此模型对测试样本进行预测，求得聚类数量 2～5 情况下的 RMSE 如表 5 所示。

根据表 5 结果，最终选择聚类数量为 5，计算结果如表 6 所示。

表 5　2 号观测点不同聚类数的均方误差

Table 5　Prediction accuracy with different cluster number on No. 2

聚类数	2	3	4	5
RMSE	1.186 3	1.434 8	1.303 7	1.133 4

表 6　2 号观测点实际噪声值和预测值　dB

Table 6　True value and prediction result on No. 2

日期	实际值	聚类再回归	ε-SVR	ARMA
2008/06/24	52.94	53.17	52.89	52.98
2008/06/25	52.23	52.77	52.43	52.87
2008/06/26	53.84	53.09	53.09	52.19
2008/06/27	51.22	51.66	51.17	52.73
2008/06/28	55.06	52.80	52.07	52.55
2008/06/29	53.20	52.79	52.59	52.95
2008/06/30	54.54	52.87	52.52	53.67

ARMA 模型是现代时间序列分析中最为常用的模型之一[7-8]，在科学研究和工程系统中具有广泛的运用，所以本文对聚类再回归、直接支持向量回归和 ARMA 模型进行比较，各模型预测精度比较如表 7 所示。

表 7　2 号观测点，各模型预测精度比较

Table 7　Prediction accuracy of different models on No. 2

模型	平均绝对百分误差 MAPE	希尔不等系数 Theil IC	均方根误差 RMSE
聚类再回归	1.536 8	0.010 7	1.133 4
ε-SVR	1.747 4	0.013 5	1.412 6
ARMA	13.888 8	0.012 6	1.333 5

对 12 号观测点，不同聚类数量的 RMSE 如表 8 所示。

表 8　12 号观测点不同聚类数的均方误差

Table 8　Prediction accuracy with different cluster number on No. 12

聚类数	2	3	4	5
RMSE	1.124 0	1.840 4	1.697 7	1.565 8

根据表 8 结果，选取聚类数量为 2，计算结果如表 9，10 所示。

表 9　12 号观测点实际噪声值和预测值　dB

Table 9　True value and prediction result on No. 12

日期	实际值	聚类再回归	ε-SVR	ARMA
2008/06/24	54.44	53.41	52.89	52.95
2008/06/25	51.65	52.20	52.43	52.47
2008/06/26	53.29	51.54	53.09	52.86
2008/06/27	50.09	49.68	51.17	52.64
2008/06/28	50.95	51.00	52.07	53.06
2008/06/29	52.24	53.10	52.59	52.90
2008/06/30	50.00	51.87	52.52	52.71

表 10　12 号观测点，各模型精度比较

Table 10　Prediction accuracy of different models on No. 12

模型	平均绝对百分误差 MAPE	希尔不等系数 Theil IC	均方根误差 MSE
聚类再回归	1.796 3	0.010 8	1.124 0
ε-SVR	2.612 3	0.014 6	1.513 3
ARMA	21.049 9	0.016 7	1.764 4

本文中所用公式如下

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\bar{y}_i-y}{y_i}\times 100\right| \tag{5}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{6}$$

$$\text{Theil IC}=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2}} \tag{7}$$

式中 y_i 和 $\hat{y}_i$ 分别为实际值和预测值。

3.3 模型比较结果

(1)平均绝对百分误差比较。一般认为 MAPE 的值低于 10,则预测精度较高,从表 6 和表 8 可以看出,先聚类再回归模型得到的 MAPE 值最小且小于 10,说明模型的预测精度最高。

(2)希尔不等系数比较[9]。希尔不等系数一般介于 0 到 1 之间,数值越小表明拟合值与真实值的差异越小,即预测精度越高。从表 6 和表 8 看出,先聚类再回归模型的希尔不等系数值远远小于 1,且最小,说明此模型的预测准确度最好。

(3)均方根误差比较。从表 6 和表 8 比较可以看出,先聚类再回归模型的均方根误差明显小于其他模型。

4 结束语

由于影响机场噪音的因素较多,如天气、节假日、飞机型号、飞行程序等,因此在机场噪声预测中引入聚类分析,先对样本进行聚类,再对同类别样本进行支持向量回归的算法,本文对先聚类再回归的算法进行了仿真实验,实验说明:将样本首先进行聚类分析,再利用相似样本进行回归分析,可以有效提高预测精度。再利用这一算法对北京某机场的实测数据进行预测,并通过绝对数值分析和相对数值分析,与直接支持向量回归模型以及现有预测方法进行比较,发现其常用的三种评价指标都符合标准,且小于其他模型,说明本文使用的基于 SVM 的先聚类再回归算法模型,预测精度最高。本文提出算法,适用场合为时间序列的单步预测,可以不断将最新样本数据加入训练模型并更新类中心点,建立动态预测模型,使模型能快速适应问题的变化,以获得更实时的预测数据。

参考文献:

[1] 周宁. 机场噪声预测与控制技术研究[D]. 杭州:浙江大学,2002.
Zhou Ning. Research on prediction and control technology of airport noise[D]. Hangzhou: Zhejiang University, 2002.

[2] 邓乃扬,田英杰. 数据挖掘中的新方法:支持向量机[M]. 北京:科学出版社,2004.
Deng Naiyang, Tian Yingjie. New method in data mining: Support vector machine[M]. Beijing: Science Press, 2004.

[3] 王海燕,卢山. 非线性时间序列分析及其应用[M]. 北京:科学出版社,2006.
Wang Haiyan, Lu Shan. The analysis and application of nonlinear time series[M]. Beijing: Science Press, 2006.

[4] 夏国恩,曾绍华,金炜东. 支持向量回归机在铁路客运量时间序列预测中的应用[J]. 计算机应用研究,2006,10:180-182.
Xia Guoen, Zeng Shaohua, Jin Weidong. Application of support vector regression in prediction of railway passenger volume time serial[J]. Application Research of Computers, 2006,10:180-182.

[5] 黄兵,郭继昌. 基于 Gabor 小波与 LBP 直方图序列的人脸年龄估计[J]. 数据采集与处理,2012,5:340-345.
Huang Bing, Guo Jichang. Age estimation of facial images based on Gabor wavelet and histogram sequence of LBP[J]. Journal of Data Acquisition and Processing, 2012,5:340-345.

[6] 董毅,程伟,张燕平,等. 基于 SVM 的先分类再回归方法及其在产量预测中的应用[J]. 计算机应用,2010,39(9):2310-2313.
Dongyi, Chengwei, Zhang Yanping, et al. Regression method based on SVM classification and its application in production forecast[J]. Journal of Computer Applications, 2010,39(9):2310-2313.

[7] Chang C C, Lin C J. LIBSVM: a library for support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[8] Chen Bojuen, Chang Mingwei, Lin Chihjen. Load forecasting using support vector machines: a study on EUNITE competition[R]. IEEE Transactions on Power Systems, 2001,19(4):1821-1830.

[9] Klaus -Robert, Müller Alex, J Smola, et al. Predicting time series with support vector machines[C]// ICANN '97 Proceedings of the 7th International Conference on Artificial Neural Networks. Berlin Heidelberg: Springer, 1997:999-1004.

作者简介:夏利(1988-),男,硕士研究生,研究方向:数据挖掘,E-mail:xialia8@126.com;王建东(1945-),男,教授,博士生导师,研究方向:数据挖掘,机器学习与知识工程;张霞(1981-),女,博士研究生,研究方向:数据挖掘;王丽娜(1979-),女,博士,研究方向:数据挖掘。