感染性腹泻周发病例数的 PCA-SVM 回归预测研究

霍 静[1] 王永明[2] 顾君忠[2]

[1](天水师范学院电子信息与电器工程学院 甘肃 天水 741001)

[2](华东师范大学计算机应用研究所 上海 200062)

摘 要 提出一个使用 PCA-SVM 进行感染性腹泻周发病例数回归预测方法,有效避免了 BP 神经网络模型存在局部极值、多重共线性的问题。以上海市 2005 年至 2008 年感染性腹泻周发病例数为样本,建立 PCA-SVM 回归模型。首先用 PCA 从统计气象因子中提取气象主成分因子,去除预报因子多重共线性,得到最终模型的解释变量,其次采用 SVM 方法构建上海市感染性腹泻周发病例数预测模型。为了说明该模型有更佳的预测效果,与 BP 神经网络模型比较拟合及预测结果。数据结果显示 PCA-SVM 回归模型预测的平均相对误差 MAPE、均方误差平方根 RMSE(数值分别为 0.2694,33.113)均小于 BP 神经网络(数值分别为 0.3745,49.909),而决定系数 R^2(数值为 0.9089)较 BP 神经网络(数值为 0.8590)更趋近于 1。证明 PCA-SVM 回归模型在感染性腹泻周发病例数预测中具有较高的预测精度和较强的泛化能力,模型对于感染性腹泻周发病例数的预测可靠,对于向公众发布腹泻预报有更好的实用价值。

关键词 PCA SVM 回归 感染性腹泻 气象资料

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2016.02.012

RESEARCH ON PCA-SVM REGRESSIVE PREDICTION OF WEEKLY CASES OF INFECTIOUS DIARRHEA

Huo Jing[1] Wang Yongming[2] Gu Junzhong[2]

[1](*School of Electronic Information and Electronical Engineering, Tianshui Normal University, Tianshui 741001, Gansu, China*)

[2](*Institute of Computer Applications, East China Normal University, Shanghai 200062, China*)

Abstract We proposed a regressive prediction method for the weekly cases number of infectious diarrhea using PCA-SVM, which effectively avoids some defects of the BP neural network model like local extremum, multicollinearity. With the weekly cases of infectious diarrhea in Shanghai from the year 2005 to 2008 being the samples, we built the PCA-SVM regressive model. First, we employed PCA to extract meteorological main principal factors from the statistical meteorological factors and removed the multicollinearity from the predictive factors, derived the explanatory variable of the final model. Secondly, we used SVM regression to build the predictive model for weekly cases number of infectious diarrhea in Shanghai. To illustrate the better prediction effect of the model, we compared it with BP neural network model in terms of fitting and prediction results. Numerical results showed that the MAPE and RMSE (0.2694 and 33.113 respectively) predicted by PCA-SVM regression model were all less than those of BP neural network model (0.3745 and 49.909 respectively). Meanwhile, its determination parameter R^2(0.9089) was further approaching 1 than that of BP neural network (0.8590). As a result, it is demonstrated in this paper that the PCA-SVM regressive model has higher prediction accuracy and stronger generalisation capability in predicting weekly cases number of infectious diarrhea, the prediction of the model is reliable on the weekly cases number of the disease, and has better practical value in publicising the diarrhea prediction.

Keywords PCA SVM regression Infectious diarrhea Meteorological data

0 引 言

全球每年约有 30 亿~50 亿人发生感染性腹泻,死亡人数约为 300 万[1,2]。研究表明,感染性腹泻的发生、流行与气象因素密切相关[3-5]。感染性腹泻一旦病发,由于其具传染性,会出现流行面广、发病率快的特点。因此探讨有效、准确的预测方法对感染性腹泻的预防控具有重要意义。

目前关于传染病预测主要有三种方法:传染病传播动力学模型[6],考虑影响传染病发病的因素很多,需要详尽的物理和气象数据,这些数据不容易获得;传统的统计模型[7,8],其中线性回归建模是最常用的方法,但对疾病建模非线性问题的预测能力并不好;智能计算技术建模,如 BPNN 神经网络、支持向量机 SVM 等。

收稿日期:2014-12-25。上海市国际科技合作基金项目(13430710100);甘肃省科技计划资助项目(1506RJZE115);甘肃省高等学校科研项目(2015B-104)。**霍静**,讲师,主研领域:数据挖掘,信息检索。**王永明**,博士。**顾君忠**,教授。

智能计算技术建模方法中,SVM 方法已在手写体识别、图像处理、信号处理等应用研究方面取得了显著成果,但在非线性特征十分显著的疾病气象预测领域的应用至今却很少[9]。截止 2014 年 12 月,以主题"SVM"在中国知网搜索相关文献,共有文献 1 079 篇,追加主题"疾病预测"后。检索结果文献仅为 19 篇。

使用上海市 2005 至 2008 年感染性腹泻周发病数和同期气象资料建立智能计算 PCA-SVM 模型,探讨 PCA-SVM 在感染性腹泻疾病预测中的可行性。同时与传统 BP 网络模型做对比,进一步验证 PCA-SVM 回归模型在腹泻发病例数预测方面的准确性,对于向公众发布腹泻预报有更好的实用价值。

1 方 法

1.1 支持向量机 SVM

支持向量机 SVM 是 Vapnik 提出的一种在模式识别与机器学习领域中的工具。主要研究在有限数据集的情况下基于数据的机器学习问题,可用于模式分类和非线性回归[10]。支持向量机主要思想是通过预先设定的非线性映射将输入空间的特征向量映射到高维特征空间,建立一个分类超平面作为决策曲面,使得正反例之间的隔离边缘被最大化,避免了在原输入空间中进行非线性曲面分割计算[11]。

(1) SVM 体系结构

SVM 体系结构如图 1 所示,其中 $x_i(i = 1,2,\cdots,n)$ 是输入变量,$K(x,x_i)$ 为核函数。常用核函数有线性核函数、多项式核函数、径向基(RBF)核函数、两层感知器核函数等。

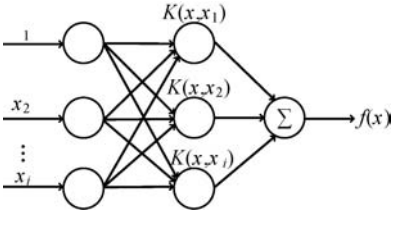

图 1 SVM 体系结构

核函数的选择是 SVM 理论的核心问题。迄今尚没有针对具体问题可以直接构造出最为适合的核函数的完备理论。其中 RBF 核属于非线性映射的核函数,可处理非线性可分情况,因而 RBF 核通常被优先考虑[12]。

(2) 算法描述及实现[13]

设给定数据集 $H = \{(x_i,y_i)\}, i = 1,2,\cdots,n$,其中 x_i 是输入变量,y_i 是期望输出值,回归估计问题就是寻找该数据集的回归(逼近)函数:

$$f(x) = w\varphi(x) + b \tag{1}$$

式中,$\varphi(x)$ 是从输入空间到高维特征空间的非线性映射,b 是偏移系数。

引入一个松弛变量 ξ_i,度量对约束条件的违反情况并采用结构风险最小化原则,将问题转化为找最小值问题:

$$\min \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{n} \xi_i$$

式中,w 是权向量,C 是惩罚参数。由于实际应用中大多数问题线性不可分,故引入满足 Mercer 条件的函数 $\varphi(x_i)$,将输入空间映射到一个可分的或者近似可分的高维的特征空间。然后在特征空间中,通过二次型寻优得到基于 SVM 的回归模型:

$$f(x) = w\varphi(x) + b = \sum_{i=1}^{n}(a_i - a_i^*)\varphi(x_i)\cdot\varphi(x_j) + b \tag{2}$$

式中,$\varphi(x_i)\cdot\varphi(x_j)$ 是向量内积运算。用核函数代替内积运算后,拟合函数为:

$$f(x) = \sum_{i=1}^{n}(a_i - a_i^*)K(x,x_i) + b \tag{3}$$

1.2 主成分分析法(PCA)

数据处理过程中统计数据经常是高维且彼此间存在一定的相关性,这些高维数据所包含的信息在一定程度上有所重叠(冗余)。主成分分析法可以很好地去除这种多重共线性,减少数据维数。

PCA 将多个变量经过线性的组合从而得出比较少的几个重要的变量的方法称为主成分分析法[14]。基本思想是提取出多维数据的主要特征(主分量),保留数据集的对方差贡献最大的特征,去掉数据相关性,在一个低维空间来快速处理数据。

1.3 模型拟合检验评价指标

评价模型拟合和外推预测效果的常用评价指标有平均相对误差(MAPE)、均方误差平方根(RMSE),决定系数 R^2,计算公式如下[15]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(x_i - \hat{x_i})}{x_i}\right|$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(x_i + \hat{x})^2}{n}}$$

$$R^2 = 1 - \frac{\sum(x_i - \hat{x_i})^2}{\sum(x_i - \bar{x})^2}$$

其中,n 为样本个数,x_i 为样本实际值,$\hat{x_i}$ 为样本预测值。*MAPE*、*RMSE* 值越小,R^2 值越接近 1,说明模型预测准确度越高。

2 基于 PCA-SVM 腹泻周发病例预测

2.1 实验资料和仿真平台

从国家疾病监测信息报告管理系统中获取 2005 年 1 月 1 日至 2008 年 12 月 31 日临床诊断或实验室确诊上海市感染性腹泻日发病数据并计算出周感染性腹泻发病例数。同期上海地区气象资料则由上海市气象局城市环境气象中心提供,有最高温度(℃)、最低温度(℃)、周平均温度(℃)、最低相对湿度(%)、平均相对湿度(%)、平均气压(hPa)、降雨量(mm)、平均日照时数(hr)、平均风速(m/s)共 9 个指标。这里 2005 至 2007 年共 157 对数据作为训练样本集,2008 年共 52 对数据作为测试数据集。试验平台采用 Matlab R2013a,结合 libsvm 工具包。

2.2 主成分提取 PCA

收集数据集属性值数量级差别很大,绝对值最小 0(降雨量),最大值 1039(日平均气压),模型采用的核函数要做向量内积运算,很容易导致计算复杂,训练时间较长,甚至会导致模型有很大的预测误差,因此,首先将训练样本和测试样本属性值用 mapminmax 函数进行归一化至 0 ~ 1。然后求出 r 矩阵。

气象属性 x_1、x_2、x_3 有很强正相关性,与气象属性 x_7 有很强负相关性,见表 1 所示。提示用 PCA 去除多重共线性,减少冗余。

表 1　r 矩阵

气象因素	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X1	1.000	0.700	0.743	0.730	0.132	0.099	-0.572	0.265	0.048
X2	0.700	1.000	0.974	0.992	0.112	0.094	-0.912	0.400	0.217
X3	0.743	0.974	1.000	0.994	0.282	0.240	-0.899	0.273	0.259
X4	0.730	0.992	0.994	1.000	0.197	0.165	-0.907	0.337	0.234
X5	0.132	0.112	0.282	0.197	1.000	0.924	-0.213	-0.546	0.292
X6	0.099	0.094	0.240	0.165	0.924	1.000	-0.222	-0.552	0.150
X7	-0.572	-0.912	-0.899	-0.907	-0.213	-0.222	1.000	-0.255	-0.304
X8	0.265	0.400	0.273	0.337	-0.546	-0.552	-0.255	1.000	0.138
X9	0.048	0.217	0.259	0.234	0.292	0.150	-0.304	0.138	1.000

计算矩阵 r 的特征值、主成分的方差贡献率、累积贡献率见表 2 所示,进而提取主成分。从表 2 中可以看到前 3 个主成分包含原来 4 个指标全部信息的 96.51%,故选作网络输入(预测因子)。

表 2　各主成分的特征值和方差贡献率

气象因素	初始特征值			提取求和的平方载荷		
	特征值	各因素方差贡献(%)	累计方差贡献率(%)	特征值	各因素方差贡献(%)	累计方差贡献率
X_1	327.4930	62.85	62.85	327.4930	62.85	62.85
X_2	158.1560	30.25	93.21	158.1560	30.25	93.21
X_3	17.1913	3.3	96.51	17.1913	3.3	96.51

2.3　训练函数选择和网络参数设置

选用径向基函数做为 SVM 回归预测模型的核函数,形式为:

$$K(x,x_i) = \exp(-\gamma \| x - x_i \|^2) \quad \gamma > 0 \qquad (4)$$

式中,x_i 是输入向量,x 是待预报因子向量,γ 是核参数,大于 0。根据式(3),选择径向基函数做为 SVM 回归预测模型的核函数后,进而最终回归函数形式为:

$$f(x) = \sum_{i=1}^{n} (a_i - a_i^*) \exp(-\gamma \| x = x_i \|^2) + b \qquad (5)$$

随参数值选取的不同,函数形态会发生相应的变化,进而引起 SVM 模型的变化。SVM 参数的选择,国际上还没有形成一个统一的模式。最优 SVM 参数的选择,目前常用的做法有交叉验证与网格搜索法进行参数优化选择[16]。这里基于 matlab 平台使用 libsvm 工具包,采用 5 则交叉验证,在反复试验的基础上确定惩罚系数 $C = 2, g = 0.5$,可以取得很好的预测结果。

2.4　实验结果分析

(1) 模型拟合检验

以 2005 年至 2007 年周气象数据和同期感染性腹泻周发病例数对预测模型进行拟合效果检验。取 2008 年的独立样本数据作为测试样本数据对模型进行外推能力检验。其中训练样本和测试样本的 R^2 分别为 0.9169 和 0.9089,说明拟合程度较好,见表 3 所示。

表 3　PCA-SVM 预测训练、测试样本性能指标

性能指标	训练样本	测试样本
MAPE	0.2423	0.2694
RMSE	30.884	33.113
R^2	0.9169	0.9089

(2) BP 神经网络

为了检验提出模型预测效果的优劣,这里和传统 BPNN 预测模型做拟合及预测效果比较。BPBP 神经网络是一种前馈型神经网络。学习过程由信号的正向传播和反向传播两个过程组成。正向传播时,输入样本从输入层传入,经各隐含层逐层处理后传向输出层。若输出层的实际输出与期望输出不符,则转入误差的反向传播阶段,误差反传阶段是将输出误差以某种形式通过隐含层向输入层逐层反传,从而获得各层单元的误差信号。此过程一直进行到网络输出的误差减少到可接受的程度,或进行到预先设定的学习时间,或进行到预先设定的学习次数为止[17]。

用 libsvm 工具包中 newff 函数建立 BP 神经网络,采用交叉验证防止训练过程中出现过拟合。通过试错法得 BPNN 最优网络结构为 4 -4 -1,学习速率设为 0.55,目标精度 0.00001,训练次数 2000 次。

(3) 模型预测效果检验

PCA-SVM、BPNN 两种模型的预测结果和比较如表 4、表 5 所示,图 2 为清晰显示预测数值对比结果,表 4 数据以月统计形式出现,数据取整。从表中数据比较可以看出采用 PCA-SVM 得到的训练样本及测试样本的 MAPE、RMSE 均小于 BPNN 而决定系数 R^2 更接近于 1。因此认为提出的 PCA-SVM 模型较 BPNN 有更好的拟合效果及预测效果。

表 4　PCA-SVM 与 BPNN 预测 2008 年周感染性腹泻发病例数结果比较

	原始值	**PCA-SVM**	**BPNN**
2008 年 1 月	143	158	187
2008 年 2 月	142	155	176
2008 年 3 月	181	191	199
2008 年 4 月	181	212	230
2008 年 5 月	301	332	347
2008 年 6 月	404	506	510
2008 年 7 月	831	806	796
2008 年 8 月	1023	897	877
2008 年 9 月	764	703	685
2008 年 10 月	386	458	481
2008 年 11 月	434	347	327
2008 年 12 月	508	335	202

表 5　PCA-SVM 与 BPNN 拟合及预测效果比较

	训练样本			测试样本		
	MAPE	RMSE	R^2	MAPE	RMSE	R^2
PCA-SVM	0.2423	30.884	0.9169	0.2694	33.113	0.9089
BPNN	0.3561	35.727	0.8992	0.3745	49.909	0.8590

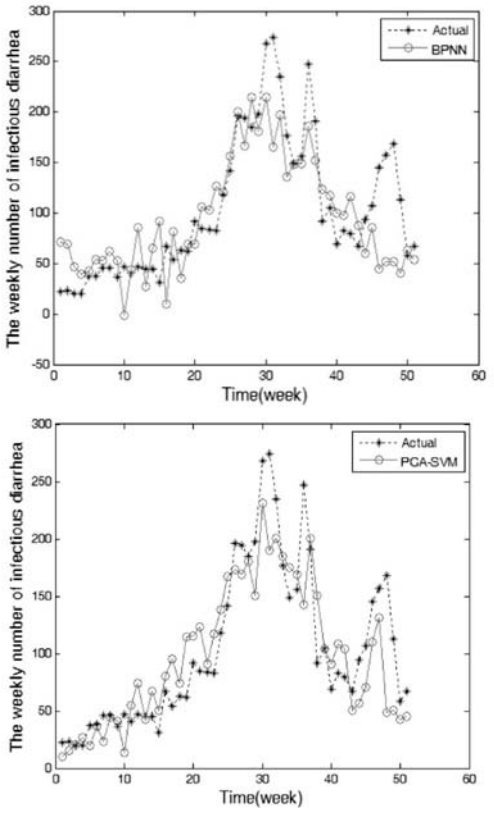

图 2 PCA-SVM、BPNN 对感染性腹泻周发病例数的预测

3 结 语

气象因素与感染性腹泻发病例数之间为非线性关系,基于 SVM 的回归预测模型可以很好地处理非线性关系。由于 BP 神经网络模型存在局部极值、多重共线的问题,提出 PCA-SVM 预测模型用于感染性腹泻周发病例数的预测模型并与 BP 神经网络模型进行比较。从表 5 实验对比结果看出,无论对训练集还是测试集, PCA-SVM 预测模型的预测结果均优于 BP 神经网络模型,比 BP 神经网络模型更适用于感染性腹泻周发病例数的预测。PCA-SVM 能够适应于多因子、多维数及样本数量有限的预测,模型泛化能力好。预测模型应用于感染性腹泻周发病例数的预测具有更高的准确度、更好的预测效果,为感染性腹泻的预测预报提供了新方法。

参 考 文 献

[1] Diarrhoeal disease. World Health Organization[EB/OL]. 2013. http://www. who. int/mediacentre/factsheets/fs330/en/.
[2] Lin M, Dong B Q. Status in epidemiological research of infectious diarrhea[J]. Chin Tropical Med,2008,8(4):675 - 677.
[3] Loyd S J, Kovats R S, Armstrong B G. Global diarrhoea morbidity, weather and climate[J]. Climate Res,2007,34(2):119.
[4] Alexander K A, Carzolio M, Goodin D, et al. Climate change is likely to worsen the public health threat of diarrheal disease in Botswana[J]. Internet Environment Res Public Health,2013,10(4):1202 - 1230.
[5] Kolstad E W, Johansson K A. Uncertainties associated with quantifying climate change impacts on human health: a case study for diarrhea[J]. Environmental Health Perspect,2011,119(3):299.
[6] 谢朝晖,黄建始. 传染病预测方法的探讨[J]. 中国全科医学,2008(1):85 - 87.
[7] Chou W C, Wu J L, Wang Y C, et al. Modeling the impact of climate variability on diarrhea-associated diseases in Taiwan[J]. Sci Total Environment,2010,409(1):43 - 51.
[8] Zhao N, Ma X H, Gan L, et al. Research on the application of Medical-meteorological foreast model of infectious diarrhea disease in Beijing[C]//IEEE Fifth International Conference,2010:149 - 156.
[9] 冯汉中,陈永义. 处理非线性分类和回归问题的一种新方法(Ⅱ)-支持向量机方法在天气预报中的应用[J]. 应用气象学报,2004,15(3):355 - 365.
[10] Vapnik V N. An overview of satistical learning theory[C]//IEEE Transactions on Neural Networks,1999,10(5):988 - 999.
[11] 杨海. SVM 核参数优化研究与应用[D]. 浙江:浙江大学电器工程学院,2014.
[12] 李阳. 多核学习 SVM 算法研究及肺结节识别[D]. 吉林:吉林大学通信工程学院,2014.
[13] 韩立群. 人工神经网络教程[M]. 北京:北京邮电大学出版社,2006.
[14] 吕建成. 神经网络中的若干问题研究[D]. 成都:电子科技大学,2006.
[15] 徐国祥. 统计预测与决策[M]. 上海:上海财经大学出版社,2008.
[16] 奉国和. SVM 分类核函数及参数选择比较[J]. 计算机工程与应用,2011(3):123 - 128.
[17] 高菡璐,兰莉,乔东菊. BP 神经网络模型用于气象因素对脑出血死亡影响的初步研究[J]. 中华流行病学杂志,2012(1):937 - 940.

(上接第 32 页)

制和总距离的最优化,对不同用户访问量进行了仿真,能科学合理地均衡数据中心的大规模访问量。整体实验结果比较合理,有一定的实用价值。

参 考 文 献

[1] 吕海燕,车晓伟,张杰. 基于“伪 HTTP Server”的 CDN 本地负载均衡实现方式[J]. 计算机技术与发展,2013,23(6):46 - 49.
[2] 赵学胜,陈军,王金庄. 基于 O-QTM 的球面 Voronoi 图的生成算法[J]. 测绘学报,2002,31(2):157 - 163.
[3] 龚咏喜,刘瑜,邬伦,等. 基于带权 Voronoi 图与地标的空间位置描述[J]. 地理与地理信息科学,2010,26(4):21 - 26.
[4] 张丽平,李松,郝忠孝. 球面上的最近邻查询方法研究[J]. 计算机工程与应用,2011,47(5):126 - 129.
[5] TechTarget 数据中心[OL]. (2013-1-25). [2014-6-10]. http://www. searchdatacenter. com. cn/shoontent_70148. htm.
[6] Nielsen F, Nock R. Hyperbolic Voronoi diagrams made easy[C]//Computational Science and Its Applications (ICCSA),2010 International Conference on. IEEE,2010:74 - 80.
[7] Bogdanov M, Devillers O, Teillaud M. Hyperbolic Delaunay complexes and Voronoi diagrams made practical[C]//Proceedings of the 29th annual symposium on Symposuim on computational geometry. ACM,2013:67 - 76.
[8] Barclay M, Galton A. Comparison of region approximation techniques based on Delaunay triangulations and Voronoi diagrams[J]. Computers Environment and Urban Systems,2008,32(4):261 - 267.
[9] Shanmugam S, Shouraboura C. Finding Optimal Allocation of Constrained Cloud Capacity Using Hyperbolic Voronoi Diagrams on the Sphere[J]. Intelligent Information Management,2012,4(5):239 - 250.
[10] 维基百科. 中华人民共和国特大城市列表[OL]. (2014-4-7).
[11] 李久刚,唐新明,刘正军,等. 基于行程距离最优及容量受限的避难所分配算法研究[J]. 测绘学报,2011(4):489 - 494.