基于百度指数的登革热疫情预测研究

王晶晶[1] 邹远强[1] 彭友松[1*] 李肯立[1] 蒋太交[1,2]
[1](湖南大学信息科学与工程学院 湖南 长沙 410082)
[2](中国科学院生物物理研究所蛋白质与多肽药物所重点实验室 北京 100101)

摘 要 基于互联网数据的传染病疫情监测成为近年来传染病防治的热点研究内容。通过对2014 年 9 月暴发的以广东省为中心的全国登革热疫情与登革热相关关键词的百度指数的关联性分析,发现地区(省、市)登革热疫情严重程度与该地区“登革热”关键词的百度指数呈很强的正相关性。为了实时地预测疫情动态,建立基于 12 个登革热相关关键词的百度指数的多元线性回归模型。在留一法交叉验证和反向测试中,该模型对于测试数据的预测值和实际值的皮尔森相关系数分别达到了 0.89 和 0.73。经实验,该预测模型能够比较准确地预测登革热疫情动态,同时该研究对于基于互联网数据的传染病疫情监测和防治具有一定的指导意义。

关键词 百度指数 登革热 定量预测模型

中图分类号 TP391 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2016.07.010

ON PREDICTION OF DENGUE EPIDEMICS BASED ON BAIDU INDEX

Wang Jingjing[1] Zou Yuanqiang[1] Peng Yousong[1*] Li Kenli[1] Jiang Taijiao[1,2]
[1](*School of Computer Science and Electronic Enginnering, Hunan University, Changsha 410082, Hunan, China*)
[2](*Key Laboratory of Protein and Peptide Pharmaceutical, National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*)

Abstract In recent years, the internet data-based epidemics surveillance for infectious diseases has been the hot topic of studies in infectious diseases prevention and treatment. Through analysing the correlation between the dengue epidemic outbreak in September, 2014 in whole China with Guangdong province as the centre and the Baidu index of the keywords correlated to dengue, we found that the severity of dengue epidemic in each province has strong positive correlation with Baidu index of keyword “dengue” in given province. For timely predicting dengue epidemic status, we built a multivariate linear regression model, which is based on the Baidu index of 12 dengue-correlated keywords. In both leave-one-out cross-validation and retrospective testing, the model performed well, with Pearson correlation coefficient between the predicted and actual epidemic size equalling to 0.89 and 0.73 respectively. It was indicated through experiment that this prediction model could be preferably accurate in predicting dengue epidemic status, at the same time our study has certain significance in terms of guidance for internet data-based surveillance, prevention and treatment of infectious diseases.

Keywords Baidu index Dengue Quantitative prediction model

0 引 言

登革热是由登革热病毒引起、伊蚊传播的一种急性传染病。临床特征为起病急骤、高热、全身肌肉、骨髓及关节痛、极度疲乏,部分患者有皮疹、出血倾向和淋巴结肿大[1]。登革热广泛流行于热带和亚热带的非洲、美洲、东南亚、西太平洋地区以及欧洲个别地区等 100 多个国家和地区。在中国,本地登革热暴发地区主要分布在广东、福建、浙江、云南和台湾,而输入性病例地区主要分布在北京、上海、香港、澳门等地[2]。如何及时有效地防治登革热已经成为了我国和世界其他多个国家和地区日益严重的公共卫生问题。

在我国,由于登革热病毒不像流感病毒那样季节性地流行,而且一直以来只是散发性流行,很少造成大的公共卫生危机。此外,登革热疫情的病例数据也很少公开。因此,目前国内针对登革热疫情监测的研究不多,特别是基于互联网数据来预测其流行动态的研究很少。2014 年 9 月在我国广东暴发了史上最大规模的登革热疫情,在短短的两个多月时间里登革热病毒感染人数超过 5 万,这对我国的社会和经济造成了很大的影响。然而此间的登革热病例数据也给我们研究基于互联网数据的传染病(尤其是登革热)疫情监测提供了一个机会。

在本文中,我们首先分析登革热在全国和广东省的疫情分布,以及研究“登革热”百度指数与地区疫情严重程度的关联

收稿日期:2015-01-28。国家自然科学基金项目(31371338);国家传染病重大专项(2013ZX10004611-002,2014ZX10004002-001);湖南大学青年教师成长计划项目(531107040720);湖南大学生物医学超算项目(531106011004)。**王晶晶**,硕士生,主研领域:生物信息学,数据挖掘。**邹远强**,博士生。**彭友松**,助理研究员。**李肯立**,教授。**蒋太交**,教授。

性,以此进一步选取与登革热相关的关键词,并分析其各关键词的百度指数与疫情动态的相关性。由此建立基于 12 个关键词的百度指数的多元线性回归模型,并将历史病例数据加入到模型训练中,通过留一法交叉验证评估模型效果,使用反向测试评价预测效果。最终我们发展了一个基于百度指数的定量预测模型来实时地预测登革热疫情的动态。

1　相关研究发展

传染病监测是预防和控制传染病疫情的核心。传统的传染病疫情监测手段主要依赖各级医疗机构、传染病预防控制中心和传染病监测哨点医院组建的监测网络提供的数据[3],整个监测体系较为完善,但存在不足。首先,数据的获取由各级单位逐层上报后汇总,会导致分析结果的滞后性;其次,该监测手段耗费大量人力物力,且病例数据很少对公众公开。而基于互联网的传染病疫情监测在很大程度上弥补了传统监测手段的不足。首先,互联网数据涵盖就诊病人和未就诊病人对传染病防控知识、疫情新闻报道等的搜索信息,数据来源的人群范围更广;其次,数据虽然集中在少数提供商手中,但其为研究用户提供了相应数据共享接口,并且数据实时公布[4]。因此,将互联网数据应用于传染病疫情的监测成为各国公共卫生研究的重要内容。

利用互联网数据监测传染病疫情的思想最先开始于 2006 年[5]。随后,各国传染病疫情监测研究者将互联网搜索引擎数据[6-11]、社交网络数据[12-15]、医疗网站数据[16]、药物销售数据[17]等应用到疫情的分析监测中。其中针对季节性流感的研究诸多,而且已经取得了很好的效果,如国外的 Ginsberg 等人[6]利用 Google 流感趋势监测流感疫情,其监测时效比 CDC 监测提前了 1 ~2 周。类似的有 Li 等人[13]利用 Twitter 数据于流感监测中,同样具有很强的实时性;在国内,李秀婷等人[7]应用 Google 搜索引擎数据研究基于互联网搜索数据的中国流感监测,从 116 个与流感相关关键词中抽取 92 个作为分析模型的搜索变量,通过交叉验证分析,最后取得了较好的模型拟合和预测效果。另袁庆玉等人[8]则是利用百度搜索引擎的百度指数数据监测中国流感趋势。针对其他传染病的研究,Milinovic 等人[9]基于 Google 搜索引擎数据利用 164 个搜索条件对 64 种传染病进行分析监测,结果显示其监测模型对其中 17 种传染病的监测效果尤为明显。这表明基于流感的监测方法对其他传染病的监测具有很大的潜在意义,尤其是对疫苗可预防、媒介传播且临床特征更明显的传染病的监测效果更好,其中包括登革热。而基于互联网数据来预测登革热也有了一些研究,影响最大的同样是来自 Google 公司的“Google Dengue Trends”。如 Althouse 等人[10]与 Chan 等人[11]应用 Google 趋势对国外登革热流行国家如新加坡等地的登革热疫情进行监测。其研究思路与“Google Flu Trends”一样,同样是选择与登革热最相关的关键词在 Google 的搜索数据,建立定量预测模型,将数据集以周为单位进行模型估计和预测,其研究取得了较好的预测效果。

由于一些原因,Google 并没有提供对于中国地区的登革热流行的预测。百度是国内市场份额最高的互联网搜索引擎[18],它推出的百度指数已经被各行各业广泛使用。在传染病监测领域,同样已经有研究使用百度指数来预测流感的流行。然而,目前还很少有使用百度指数和其他互联网数据来预测登革热的流行。

2　登革热疫情分布

2014 年 9 月,登革热在中国广东一带暴发,病例主要分布在广东、广西、云南、福建和台湾(如图 1(a)所示)。截止 10 月 31 日,全国登革热病例数超过 5 万,广东省疫情最为严重,已累计报告登革热病例 42 358 例;台湾省累计报告 7425 例;广西、云南、福建省累计报告的本地登革热病例均超过 100 例;海南、北京、湖南、浙江、澳门、香港地区累计报告的登革热病例数均在 100 例以下,而且主要是输入性病例。进一步分析广东省的登革热疫情(如图 1(b)所示),发现超过 80% 的病例(累计 35 237 例)都分布在广州,其次是佛山(累计 3411 例),其余市的病例数均在 1000 例以下。由登革热引发的死亡病例也主要分布在广州和佛山,分别有 5 例和 1 例病例死亡。

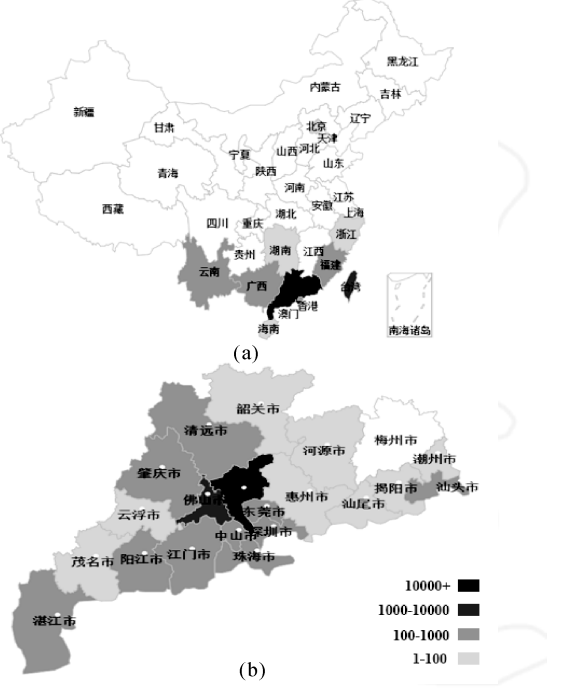


图 1　登革热疫情在全国、广东省的病例分布

3　实验数据与方法

3.1　数　据

(1) 登革热病例

本文使用的登革热病例数据来源于中国卫生与计划生育委员会官方网站、各省卫生与计划生育委员会官方网站以及网络新闻报道搜索。病例数据包括全国各疫情省份和广东省各疫情市截止 2014 年 10 月 31 日的总病例数,以及广东省从 2014 年 9 月 22 日到 2014 年 10 月 30 日间每日新增病例数,由于除广东省的其余省登革热疫情较轻缓,统一报道较少,因此结合网络新闻搜索共同取得。

(2) 百度指数

本文使用的百度指数数据来源于百度指数平台(http://index.baidu.com)。百度指数是指关键词在相应时间段内的搜索量数据。本文采集的数据集以天为单位。由于只能得到 2014 年 9 月 22 日到 2014 年 10 月 30 日间广东省的登革热每日新增病例数,因此无特别说明外,实验所使用的关键词的百度指

数都是指这段时间的数据。

3.2 方法学

(1) 关键词选取

本文根据登革热定义和临床症状等方面选取了 15 个与登革热密切相关的搜索关键词,去除未被百度指数平台收录的 3 个关键词,剩下 12 个关键词,分别是“登革热”、“伊蚊”、“皮疹”、“淋巴结肿大”、“头痛”、“恶心”、“呕吐”、“腹泻”、“便秘”、“关节痛”、“发烧”、“皮肤瘙痒”。

(2) 预测模型

$$D_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i B_{i,t} + \varepsilon_t \quad (1)$$

$$D_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i B_{i,t} + \beta_j D_{t-j} + \varepsilon_t \quad (2)$$

本文应用的模型为多元线性回归模型,在模型式(1)中,D_t 为第 t 天的登革热新增病例数,$B_{i,t}$ 表示第 i 个关键词在第 t 天的百度指数数值,n 表示模型中包含的搜索关键字的个数,$n \in [1, 12]$,ε_t 表示模型中的残差项。在模型式(2)(改进的模型)中,D_{t-j} 表示对于第 t 天向前偏移 j 天后得到的登革热每日新增病例数值,$j \in [1,7]$。

(3) 相关定义

留一法交叉验证 假设有 n 条数据,将每一条数据作为测试集,其余 $n-1$ 条数据作为训练集。重复方法使每条数据都被作为一次测试集。最后本文用测试集的预测值和实际值之间的相关性作为评价指标。

反向测试 指用过去的时间序列数据做训练集,预测未来的时间序列数据。假设数据集共 M 条数据,用后 N 条数据作测试集。以测试其中的第 n 点为例,我们将前$(M-N+n-1)$ 条数据作为训练集构建模型,预测第 n 点的值。重复方法 N 次,最后本文将预测值和实际值之间的相关性作为评价指标。

逐步回归 为建立最优回归方程,从可供选择的所有变量中选出对 D_t 有显著影响的变量建立“最优”回归方程。

(4) 统计学分析

本文的相关性分析采用皮尔森相关系数(Pearson)和斯皮尔曼相关系数(Spearman)的方法,使用 R 语言中的 cor()函数完成。多元线形回归模型使用 R 语言中的 lm()函数完成,逐步回归使用 R 语言中的 step()函数完成。预测模型的验证采用留一法交叉验证 LOOCV(Leave-one-out cross validation)和反向测试(Retrospective test),R 软件的版本为 R 3.1.2。

4 实验结果与分析

4.1 百度指数与地区疫情严重程度的相关性

为了定性地衡量百度指数与登革热疫情的关联性,我们首先分析了关键词“登革热”的百度指数与登革热疫情严重程度的相关性。表 1 展示的是在登革热流行期间(2014 年 9 月 1 日到 2014 年 10 月 31 日)各个疫情省份“登革热”的百度指数中位数,以及相应省份截至 2014 年 10 月 31 日的总病例数。我们发现整体上省份病例数越多其百度指数越高,经计算,两者存在明显的正相关:皮尔森相关系数(PCC)为 0.997,斯皮尔曼相关系数(SCC)为 0.738。

表 1 “登革热”百度指数中位数与病例总数

省份	百度指数中位数	病例总数
广东	12 916	42 358
广西	1245	573
云南	349	132
福建	1228	257
海南 *	803	5
浙江 *	1225	16
湖南 *	1067	15
北京 *	1156	23

注:* 表示输入性病例省份

进一步将关联性分析细化,对广东省内各个疫情市(20 个市)的“登革热”百度指数中位数与病例总数进行相关性分析,同样发现两者之间存在很强的相关性($PCC = 0.928$,$SCC = 0.752$),两者的关系如图 2 所示。

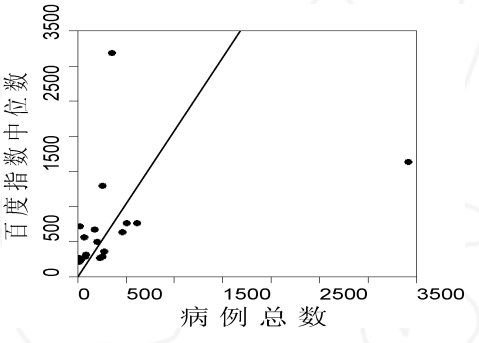
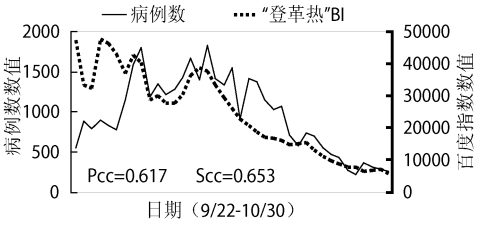


图 2 广东省各疫情市(除广州)百度指数中位数与该市的病例总数的关系

4.2 各关键词的百度指数与疫情变化的相关性

前面分析表明,从总体上来说,某地区的登革热疫情的严重程度与该地区的“登革热”百度指数相关性较强,说明可以使用百度指数来定性地评估登革热疫情的严重性。那么它是否能够用来预测登革热疫情的动态变化? 由于此次登革热疫情主要发生在广东省,因此为定量评估百度指数与疫情变化的相关性,本文针对广东省的疫情动态进行研究。除了关键词“登革热”,本文另外选择了 11 个与登革热相关的关键词,分析其在广东省范围内的每日百度指数与该省登革热每日新增病例数的相关性。图 3(X 轴日期间隔为天;Y 轴采用双坐标,左 Y 轴为广东省每日新增病例数(对应实曲线),右 Y 轴为关键词的百度指数数值(对应虚曲线);BI 为百度指数缩写)举例展示相关性较强的 5 个关键词的百度指数与病例数的曲线。经分析,登革热最常见的症状“皮疹”的百度指数与每日新增病例数的相关性最高($PCC = 0.825$,$SCC = 0.823$);此外,登革热名词“登革热”和登革热的常见症状“发烧”、“皮肤瘙痒”以及登革热的传染源“伊蚊”的百度指数都与病例数有非常强的在时间维度上的正相关。其他关键词的百度指数则与登革热病例数的相关性较弱。



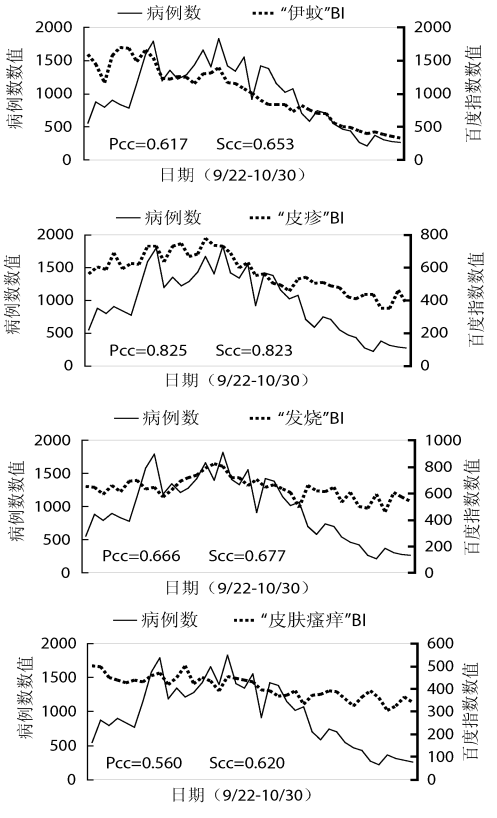


图 3　广东省每日新增病例数与各个登革热相关关键词的百度指数的关系

4.3　模型预测

为了进一步基于百度指数预测登革热疫情动态,本文重点研究基于百度指数来预测广东省的登革热疫情,建立多元线形回归模型。该模型以上面相关性分析中与登革热疫情相关的 12 个关键词的百度指数作为自变量,以广东省每日新增病例数作为因变量,该模型增加使用逐步回归方法去除回归效果不够明显的自变量。

(1) 模型训练

为了检测模型的效果,我们首先将所有数据(2014 年 9 月 22 日至 2014 年 10 月 30 日期间的广东省每日新增病例数与 12 个关键词在此期间的每日百度指数,39 * 13)作为训练集进行测试。

Input:$S=\{(C_i, X_{i\_1}, X_{i\_2}, \cdots, X_{i\_12}),\ i=1,2,\cdots,39\}$

Process:

Step1　//在训练集 S 上进行多元线性回归分析

　　$Ms <- lm(C \sim X_1 + X_2 + \cdots + X_{12}, S)$

Step2　//逐步回归

　　$Ss <- step(Ms)$

Step3　//预测值

　　$Ps <- predict(Ss, S)$

Step4　//相关性

　　cor (C, Ps[,1])

Output:$\{(C_i, Ps[n,1]),\ i, n=1,2,\cdots,39\}$ 相关系数

模型的训练效果显示,其在训练数据上的预测值和实际值两者的 *PCC* 达到了 0.874,说明模型在训练集上的效果较好。图 4(a)表示该模型在训练数据上的预测值和实际值的关系。

(2) 模型估计

进一步我们使用留一法交叉验证来评估该模型的效果,循环将 39 -1 天的数据作为训练集,其中另 1 天的数据作为测试集。

Input:$S=\{(C_i, X_{i\_1}, X_{i\_2}, \cdots, X_{i\_12}), i=1,2,\cdots,39\}$

Process:

Step1　For i = 1,2,…,39

　　//在 S 上除去第 i 天的数据得到训练集

　　$T <- S[-i,]$

　　//在训练集 T 上进行多元线性回归分析

　　$Ms <- lm(C \sim X_1 + X_2 + \cdots + X_{12}, T)$

　　//逐步回归

　　$Ss <- step(Ms)$

　　//预测值

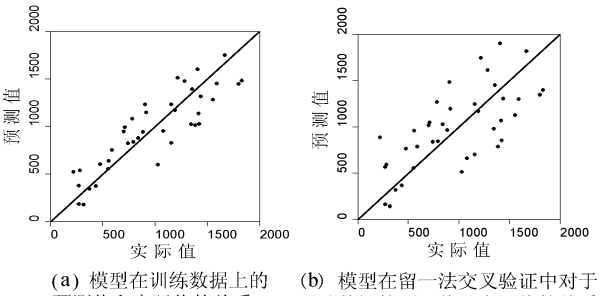
　　$Ps[i] <- predict(Ss, S)$

Step2　//相关性

　　cor (C, Ps)

Output:$\{(C_i, Ps[i]),\ i=1,2,\cdots,39\}$ 的相关系数

模型的评估效果显示,其在留一法交叉验证的测试集上模型的预测值和实际值的 *PCC* 为 0.691,说明该模型在测试数据上的效果也较好。图 4 表示模型的效果。



(a) 模型在训练数据上的预测值和实际值的关系　(b) 模型在留一法交叉验证中对于测试数据的预测值和实际值的关系

图 4　基于登革热相关关键词预测登革热疫情的模型的效果

(3) 模型预测

为了测试模型在实际的登革热疫情预测中的效果,本文对该模型做了反向测试,即用某天之前的数据训练模型。然后用得到的模型去预测该天的病例数,进而分析其预测值和实际值的相关性。在本实验中,我们使用前 31 天的数据预测后 8 天的登革热病例数。

Input:$S=\{(C_i, X_{i\_1}, X_{i\_2}, \cdots, X_{i\_12}),\ i=1,2,\cdots,39\}$

Process:

Step1　For j = 1,2,…,8

　　//取 S 的前 j +30 天的数据作为训练集

　　$T <- \{Si,\ i=1,2,\cdots,j+30\}$

　　//在训练集 T 上进行多元线性回归分析

　　$Ms <- lm(C \sim X_1 + X_2 + \cdots + X_{12}, T))$

　　//逐步回归

　　$Ss <- step(Ms)$

　　//预测值

　　$Ps[j] <- predict(Ss, Sj_{+31})$

Step2　//相关性

　　cor (C, Ps)

Output:$\{(C_i, Ps[j]),\ i, j=1,2,\cdots,8\}$ 的相关系数

通过模型预测得到后 8 天的实际值,发现该模型在反向测试中的效果较差,预测值和实际值的皮尔森相关系数只有 0.379。

4.4　改进的模型预测

考虑到历史的登革热疫情也对当前登革热疫情有一定影响,因此本文将当前登革热疫情 *N* 天(*N* = 1 ~ 7)前的登革热病例数也作为变量加到定量预测模型中,然后评估新模型的效果。

以反向测试举例说明新模型的预测算法：

Input：$S = \{(C_i, X_{i\_p}, X_{i\_1}, X_{i\_2}, \cdots, X_{i\_12}), i = 1,2,\cdots,39\}$

Process：

Step1 For N = 1,2,…,7

For j = 1,2,…,8

//取 S 偏移 N 天后的前 j + 30 - N 天的数据为训练集

T <- {Si, i = 1,2,…,j + 30 - N}

//在 T 上进行多元线性回归分析

Ms <- lm(C ~ Xp + X_1 + X_2 + … + X_{12}, T)

//逐步回归

Ss <- step(Ms)

//预测值

Ps[j] <- predict(Ss, Sj_{+31-N})

Step2 //相关性

cor (C, Ps)

Output：偏移 1 ~ 7 天的相关系数集 Cor[i], i = 1,2,…,7。

表 2 展示了分别把 1 ~ 7 天前的历史登革热病例数作为变量增加到模型中得到的新模型在留一法交叉验证中的效果。可以发现，整合历史数据之后，模型不管是在留一法交叉验证还是反向测试中的效果明显增加，其中在留一法交叉验证中，其预测值与实际值的 *PCC* 均在 0.75 以上；在反向测试中，预测值与实际值的 *PCC* 最高达到了 0.733。

表 2 不同偏移时间的模型留一法交叉验证和反向测试效果

N	留一法交叉验证	反向测试
7	0.885	0.733
6	0.915	0.585
5	0.846	0.273
4	0.775	0.559
3	0.803	0.534
2	0.798	0.027
1	0.864	0.233

图 5 表示在整合 7 天前的历史数据时模型在留一法交叉验证和反向测试中的预测值和实际值的关系。从图 5(a)可以看到在留一法交叉验证中，整合 7 天前的历史数据使得测试值与实际值更为接近；从图 5(b)可以看到在反向测试中，整合 7 天前的历史数据使得测试值与实际值不仅相关性较强，而且比较接近。因此加入历史登革热病例数据到模型训练中使得模型的预测效果得到了很大的提高。

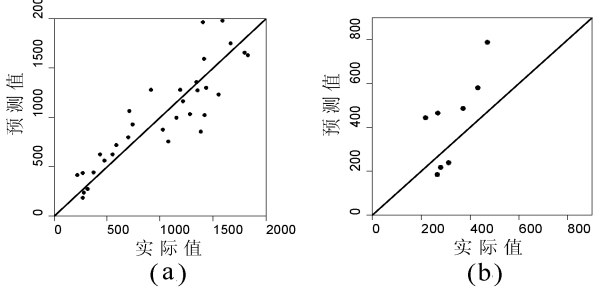


图 5 整合 7 天前的历史登革热病例数据得到的改进的模型在留一法交叉验证(a)和反向测试(b)中其预测值和实际值的关系

5 结 语

本文通过对登革热相关关键词的百度指数与实际登革热疫情进行相关性分析，发现地区登革热疫情的严重程度与该地区的百度指数存在很强的关联性。与此同时，在广东省登革热暴发期间，每日的登革热新增病例数与登革热相关关键词的百度指数也存在明显的正相关。分析发现，与登革热相关的几个关键词，如“登革热”、“皮疹”、“发热”、“伊蚊”等的百度指数与实际的登革热疫情之间存在较强的正相关。基于与登革热相关的 12 个关键词的百度指数建立的登革热预测模型在留一法交叉验证和反向测试中的效果也较好。因此本文构建的定量预测模型能够比较准确地预测广东省的登革热疫情动态。

由于此次登革热在广东省暴发持续的时间较短，因此本研究的一个不足之处在于研究的时间段不长。然而，本研究发现的登革热相关关键词的百度指数和登革热疫情的关联性非常明显，而且基于它们建立的模型也确实能够较为准确地预测登革热的实时疫情。因此，本研究对于国内使用互联网数据监测传染病(特别是登革热)的工作具有一定的参考价值和指导意义。

参 考 文 献

[1] 中国疾病预防控制中心[EB/OL]. (2014-11-06). [2015-01-23]. http://www.china.cdc/gwxx/201411/t20141106_10630.htm.

[2] 何剑峰. 登革热流行趋势及防控策略[J]. 实用医学杂志，2014(19)：3462 - 3463.

[3] 突发公共卫生事件与传染病疫情监测信息报告管理办法(卫生部令第 37 号，2006 年 8 月修改版)[EB/OL]. (2009-01). [2015-01-23]. http://www.nhfpc.gov.cn/jkj/s7913/200901/896c7b47c2d84b8b84586f17ade28d71.shtml.

[4] 李锐，王增亮，张志杰. 互联网搜索数据与流感预警[J]. 中华流行病学杂志，2013(1)：101 - 103.

[5] Eysenbach G. Tracking flu-related searches on the web for syndromic surveillance[J]. AMIA Annu Symp Proc，2006(1)：244 - 248.

[6] Ginsberg J，Mohebbi M H，Patel R S，et al. Detecting influenza epidemics using search engine query data[J]. Nature，2009，457(7232)：1012 - 1014.

[7] 李秀婷，刘凡，董纪昌，等. 基于互联网搜索数据的中国流感监测[J]. 系统工程理论与实践，2013(12)：3028 - 3034.

[8] Yuan Q Y，Nsoesie E O，Lv B，et al. Monitoring influenza epidemics in china with search query from Baidu[J]. PloS ONE，2013，8(5)：1 - 7.

[9] Milinovich G J，Avril S M，Clements A C，et al. Using internet search queries for infectious disease surveillance：screening diseases for suitability[J]. BMC Infectious Diseases，2014，14(1)：3840.

[10] Althouse B M，Ng Y Y，Cummings D A T. Prediction of Dengue Incidence Using Search Query Surveillance[J]. PloS Neglected Tropical Diseases，2011，5(8)：e1258.

[11] Chan E H，Sahai V，Conrad C，et al. Using Web search Query Data to Monitor Dengue Epidemics：A New Model for Neglected Tropical Disease Surveillance[J]. PloS Neglected Tropical Diseases，2011，5(5)：e1206.

[12] Gu H，Chen B，Zhu H，et al. Importance of Internet Surveillance in Public Health Emergency Control and Prevention Evidence From a Digital Epidemiologic Study During Avian Influenza A H7N9 Outbreaks[J]. J Med Internet Res，2014，16(1)：e20.

[13] Li J，Cardie C. Early Stage Influenza Detection from Twitter[J]. Eprint arXiv，2013.

[14] Signorini A，Segre A M，Polgreen P M. The use of Twitter to Track Levels of Disease Activity and Public Concern in the U. S during the Influenza A H1N1 Pandemic[J]. PLoS ONE，2011，6(5)：e19467.

(下转第 78 页)

验证(LOO-CV)的实验表现。

表 2 不同数据留一交叉验证(LOO-CV)结果(%)

数据集	原特征 + SVM	Wrapper + SVM	本文方法 + SVM
Leukemia	98.61	100.00	100.00
Gliomas	80.00	96.00	98.00
DLBCL	97.40	100.00	100.00
Colon	82.26	95.16	98.39

将样本按根据表 3 的方法划分为训练集和测试集,在所有样本中选择 1/3 做测试集,余下 2/3 做训练集,保证训练集和测试集中两类样本数量的比例大致相同。

表 3 样本训练集和测试集划分方法

数据集	训练集		测试集	
	Class 1	Class 2	Class 1	Class 2
Leukemia	31	16	16	9
Gliomas	18	14	10	8
DLBCL	38	12	20	7
Colon	14	26	8	14

将表 1 中的四个数据集的样本按照表 3 的方法划分训练集和测试集,然后分别用这三种方法测试,实验得到的准确率(Accuracy)和 F 指标(F1-Score)[13] 见表 4 所示。

表 4 样本按表 3 划分训练集和测试集的实验结果(%)

数据集	原特征 + SVM		Wrapper + SVM		本文方法 + SVM	
	准确率	F 指标	准确率	F 指标	准确率	F 指标
Leukemia	97.24	95.92	99.00	98.53	99.20	98.57
Gliomas	78.56	80.67	95.40	95.60	94.61	94.87
DLBCL	96.52	93.07	95.24	91.25	98.85	97.72
Colon	81.23	73.16	91.32	87.49	95.27	93.26

从表 2 可以看出,用三种方法分别对这四个数据集进行留一法交叉验证,本文方法的表现优于 Wrapper 方法和原始特征分类方法,并且该实验结果不存在随机性,证明了本文方法实验表现效果较好。

从表 4 可以看出,对于数据集 Leukemia,DLBCL 和 Colon,本文方法的分类准确率和 F 指标较高,对于数据集 Gliomas,本文方法的分类准确率和 F 指标略低于 Wrapper 方法。因此,综合考虑表 2 和表 4 的实验结果可以看出,本文方法在不同的数据集上表现都同样稳定,而且分类的准确率也比较高,从而证明该方法的有效性。

3 结 语

DNA 微阵列数据为肿瘤疾病的诊断开辟了新的思路,受实验环境和实验成本等因素的限制,DNA 微阵列数据普遍含有噪声数据和冗余基因,而且具有高维、小样本等特点,这些特点使得传统的机器学习算法无法在微阵列数据上发挥高效的作用。本文提出了一种混合型的特征选择的方法,并将该方法应用于高维肿瘤 DNA 微阵列数据的分类,对于高维的肿瘤基因数据,基因之间必然存在冗余性和不相关性,剔除冗余基因能大大降低矩阵的维数。本文方法在剔除冗余基因的时候独创性地考虑了样本的标签,综合分析多类相关矩阵以剔除冗余特征,最终通过评价函数筛选得到最优特征子集。通过实验结果可以看出,本文提出的方法是有价值的。

参 考 文 献

[1] 李波. 基于流形学习的特征提取方法及其应用研究[D]. 安徽:中国科学技术大学,2008.

[2] 王娟,慈林林,姚康泽. 特征选择方法综述[J]. 计算机工程与科学,2005,27(12):68-71.

[3] 张琳,陈燕,李桃迎. 决策树分类算法研究[J]. 计算机工程,2011,37(13):66-70.

[4] 张翔,邓赵红,王士同. 极大熵 Relief 特征加权[J]. 计算机研究与发展,2011,48(6):1038-1048.

[5] Zhang F P,Qiu Z G,Feng X T. Non-complete Relief Method for Measuring Surface Stresses in Surrounding Rocks[J]. J. Cent. South Univ,2014,21(9):3665-3673.

[6] 范文兵,王全全,雷天友. 基于 Q-relief 的图像特征选择算法[J]. 计算机应用,2011,31(3):724-728.

[7] Nutt C L,Mani D R,Betensky R A. Gene Expression-Based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification[J]. Cancer Res,2003,63(7):1602-1607.

[8] 鲜晓东,樊宇星. 基于 Fisher 比的梅尔倒谱系数混合特征提取方法[J]. 计算机应用,2014,34(2):558-561.

[9] 章舜仲,王树梅. 相关系数矩阵与多元线性相关分析[J]. 大学数学,2011,27(2):195-198.

[10] Golub T R,Slonim D K,Tamayo P. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring[J]. Science,1999,286(15):531-537.

[11] Alizadeh A A,Eisen M B,Davis R E. Distinct types of diffuse large B-cell lymphoma identified by gene expression pmrdillg[J]. Nature,2000,403(6769):503-511.

[12] Alon U,Barkai N,Notterman D A. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays[J]. Proc Natl Acad Sci USA,1999,96(12):6745-6750.

[13] 刘诚. 蛋白质相互作用界面中热点残基预测方法的研究[D]. 湖北:武汉科技大学计算机科学与技术学院,2012.

(上接第 46 页)

[15] Fung I C,Fu K W,Ying Y C,et al. Chinese social media reaction to the MERS-CoV and avian influenza A(H7N9) outbreaks[J]. Infectious Diseases of Poverty,2013,2(1):31.

[16] Hulth A,Rydevik G,Linde A. Web Queries as a Source for Syndromic Surveillance[J]. PLoS ONE,2009,4(2):e4378.

[17] Pivette M,Mueller J E,Crepey P,et al. Drug sales data analysis for outbreak detection of infectious diseases: a systematic literature review[J]. BMC Infectious Diseases,2014,14(1):604.

[18] 中国互联网络发展状况统计报告[EB/OL]. (2014-01). [2015-01-23]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201403/P020140305346585959798.pdf.