

· 论 著 ·

使用 R 语言扩展中国疾病预防控制中心信息系统预测重庆市结核病疫情*

汪 洋¹, 胡代玉², 刘 璵², 王润华³, 杨德香³, 易 静^{3△}

(1. 重庆医科大学实验教学管理中心公共卫生实验教学中心, 重庆 401331; 2. 重庆市结核病防治所, 重庆 400050; 3. 重庆医科大学公共卫生与管理学院卫生统计学教研室, 重庆 400016)

摘要:目的 扩展《中国疾病预防控制中心信息系统》, 实现结核病发病的在线预测功能, 为将来的疾病预防控制中心工作的资源调配及优化起到辅助决策支撑作用。方法 采用离线的《中国疾病预防控制中心信息系统》中重庆市 2005 年至 2009 年每个月的结核病发病登记数据, 以 Java 平台为基础运行环境, 采用 R 语言分析预测引擎, 应用神经网络模型、灰色模型、时间序列模型对结核病发病率进行在线预测分析。结果 灰色模型和时间序列模型对结核病预测的误差较大, 而神经网络模型对于结核病发病的预测相对误差不超过 5%。结论 使用 R 语言扩展《中国疾病预防控制中心信息系统》具有可行性, 可用于指导疾病预防控制中心工作中的资源调配及决策支撑。系统扩展后对于结核病发病可实现单机、网络的在线预测。

关键词: 编程语言; 预测; 中国疾病预防控制中心信息系统; 结核病疫情

doi: 10.3969/j.issn.1671-8348.2013.13.005

文献标识码: A

文章编号: 1671-8348(2013)13-1454-03

Research on function expanding of China Information System of Disease Control and Prevention using R language to predict the tuberculosis prevalence*Wang Yang¹, Hu Daiyu², Liu Ying², Wang Runhua³, Yang Dexiang³, Yi Jing^{3△}

(1. Laboratory Teaching Center for Public Health, Laboratory Teaching & Management Center, Chongqing Medical University, Chongqing 401331, China; 2. Chongqing Institute of Tuberculosis Prevention and Treatment, Chongqing 400050, China; 3. Teaching & Research Unit of Hygiene Statistics, School of Public Health and Management, Chongqing Medical University, Chongqing 400016, China)

Abstract: Objective To expand the function of China Information System of Disease Control and Prevention for realizing the online prediction of tuberculosis prevalence, in order to provide the support of decision making for allocation and optimization of resources in disease prevention and control. **Methods** The off-line registered data on tuberculosis morbidity in Chongqing originated from China Information System of Disease Control and Prevention during 2005—2009 were collected. Then these data were analyzed by computing engine programmed with R language, using different models such as neuro-network model, gray model and time series model to realize the online predicting of tuberculosis prevalence based on Java platform. **Results** Among these three different models, the relative error of neuro-network in predicting the TB prevalence was less than 5% and therefore was much better than the other two models in prediction. **Conclusion** The function expanding of China Information System of Disease Control and Prevention is feasible using R language. Moreover, the expanded system could make the prediction of TB prevalence available on personal computer or online. Therefore, it could be used to provide guidance and support in resources allocation and decision-making in disease control and prevention.

Key words: programming languages; forecasting; china information system of disease control and prevention; tuberculosis prevalence

《中国疾病预防控制中心信息系统》为一个覆盖中央、省、地(市)县(市区)4 级疾病预防控制中心和乡镇卫生院以上医疗单位的广域网, 连接着全国 31 个省省级疾病控制中心和 331 个地(市)级疾病控制中心的计算机局域网及其下属 2 863 个县(市、区)疾病控制机构的计算机网站。《中国疾病预防控制中心信息系统》为该网络系统最主要、也是最重要的部分^[1]。《中国疾病预防控制中心信息系统》以分布式的数据存储、统计查询和数据管理为主, 尚无结核病发病率预测功能, 本文主要引用该系统的专病——结核病数据为例说明如何使用 R 语言扩展系统, 使其具备结核病发病预测功能。根据结核病变化规律建立的结核病疫情发展预测模型, 包含了影响结核病疫情发展的众多因素, 逼近准确的结核发病率预测可以量化对结核病的控制和管理的决策, 是有效的结核病防治工作的前提。简化数据分析

与预测的难度, 降低使用门槛, 使未同时掌握数理统计与计算机专业知识的群体能够简单快捷地对数据进行分析与预测。本文所涉及软件均使用免费开源软件, 可降低科研成本。

1 R 语言与 Java 语言

R 语言是由 AT&T 贝尔实验室所创的 S 语言发展出的有着统计分析及强大作图功能的软件系统。R 语言内含的作图函数能将产生的图片展示在一个独立的窗口中, 并能将之保存为各种形式的文件(jpg、png、bmp、ps、pdf、emf、pictex、xfig; 具体形式取决于操作系统)。统计分析的结果也能被直接显示出来, 一些中间结果(如 P 值、回归系数、残差等)既可保存到专门的文件中, 也可以直接用作进一步的分析^[2-3]。与 SPSS、MATLAB 等相比, R 语言作为一个开源的免费软件, 其价值得到不断地延伸, 主要表现在: 可以跨平台运行, 对矩阵的操作强

* 基金项目: 国家自然科学基金资助项目(30872061); 重庆市科委自然科学基金资助项目(2009BB5415)。 作者简介: 汪洋(1970~), 副教授, 博士研究生, 主要从事预防医学研究。 △ 通讯作者, Tel: (023)68485259; E-mail: yijing@sina.com。

大而高速,拥有许多可用的附加包,灵活的编程环境,对于已经广泛使用的统计软件 S/S. Plus 是可兼容的^[4-5]。Java EE 是目前最为流行的主流技术之一。它是 J2EE 开发进一步简化的升级版,使得 Java 企业级应用开发变得简单和快捷,其强大的 API 降低了应用复杂性,提高了应用的性能,扩展了应用的空间^[6-8]。本文主要描述 Java 语言如何控制 R 服务,通过 R 连接器调用 R 服务提供的各种分析与预测模型,以及相关的数据处理实现。其余数据管理和 WEB 用户界面等通用的程序实现暂不赘述。

2 程序实现

2.1 程序架构 相关支撑软件均为免费开源软件,故可采用 Windows 或 Unix 系列的操作系统。数据库独立运行于操作系统之上,与程序采用 JDBC 通信。程序使用 Java 语言编写完成,运行在 Java 平台之上,主要分为连接管理(控制 R 服务的启动与停止、并发计算时资源同步调用、预测数据输入输出控制)、数据管理(根据预测需求,从数据库批量提取数据结果处理转换后提交给 R 引擎进行预测、后期数据维护以及原始数据的查询和简易统计)、业务处理(识别用户提交的各种请求,并根据请求组织图文信息显示出来)和 WEB 用户界面(以 WEB 形式提供用户操作与交互的界面),见图 1。

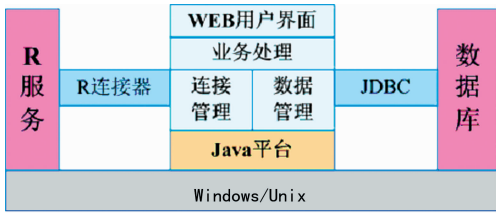


图 1 程序架构

2.2 系统流程 (1)用户进入用户界面 A(WEB 形式的用户操作界面),在 A1.上执行预测分析操作。(2)用户进入 B(业务层,内部业务处理层),根据用户要求,从数据库提取数据(B2.);将数据提交给预测分析处理模块(B3.);调用 R 服务进行预测分析(B4.),并取回计算结果;将计算结果绘制为图表,并返回给用户界面用于直观地显示(B5.)。C、D 为相对独立的外部服务,见图 2。

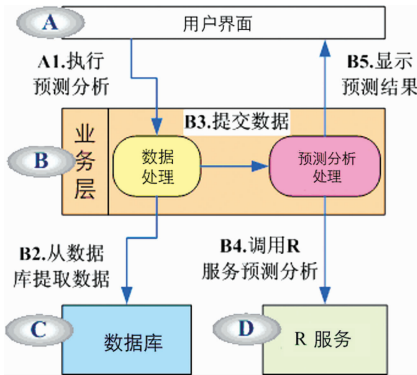


图 2 系统流程

2.3 关键程序实现

2.3.1 Java 程序调用 R 服务 (1)调用方式选择。使用 Java 程序调用 R 服务,可通过多种方法:①在 Java 中执行 R 脚本并获取输出数据,每次都需要创建一个 R 进程效率低下,可操控性能低;②使用 R 组件 Rserve,该组件基于 TCP/IP 提供 R 计算服务,只需一个进程,通过 TCP/IP 长连接,可反复调用,且可操控性能较好;③使用共享内存调用,实现较为复杂。对比

上述 3 种方式,本程序采用第 2 种实现。(2)所需开发软件包。有超过 1 800 个免费的预测分析功能包供下载,地址: <http://cran.r-project.org/web/packages>。R 语言分析预测包^[9]: AMORE、Rserve,下载地址: http://cran.csdb.cn/web/packages/available_packages_by_name.html; Java 语言程序包: REngine.jar、Rserve.jar,下载地址: <http://rosuda.org/Rserve/download.shtml>。(3)Rserve 运行环境配置(以 Windows 操作系统为例)。环境变量:R_HOME=R 语言安装路径,路径中不建议使用中文或包含空格的名称;路径:%R_HOME%\bin\i386;执行目录:%R_HOME%\library\Rserve;执行命令:Rserve.exe--RS-encoding=utf8,将语言编码设置为 utf8,中文方可正常显示。(4)调用代码。String cmd=OLStatContext.getRScriptDirectory()+"\\rserve.cmd";//Rserve 启动批处理脚本文件。cmd=cmd.replaceAll("/", "\\");started=Program.launch(cmd);//使用 SWT 组件中的 Program 工具类在单独的进程中启动 Rserve。RConnection connection=new RConnection();//通过默认端口连接 Rserve REXP x=connection.Eval("R.version.string");//执行一个字符串表达式,获取 R 服务版本 System.out.println(x.asString());//显示从 R 服务中获取的字符串值 cmd="source(\""+cmd+"\")";//初始化命令,加载 GM 函数、AMORE 程序包,仅需要在 Rserve 启动时运行一次即可。conn.voidEval(cmd);//执行外部 R 语言脚本。

2.3.2 调用时间序列模型 (1)主要调用代码:.....RConnection conn=RServerManager.getConnection();//获取 R 服务连接。conn.assign("srcData",srcData);//时间序列变量赋值。.....String cmd=OLStatContext.getRScriptDirectory()+"arima-1.r";//获取预先准备好的 r 执行脚本文件路径。cmd=cmd.replaceAll("\\", "/");//将路径中的“\”替换为“/”。cmd="source(\""+cmd+"\")";//R 调用外部脚本语法。conn.voidEval(cmd);//执行 R 语言代码,调用外部脚本。RList rlist=conn.eval("predictValues['pred']").asList();//获取计算输出值。(2)程序界面截图,见图 3~5。



图 3 时间序列模型预测数据输入界面图

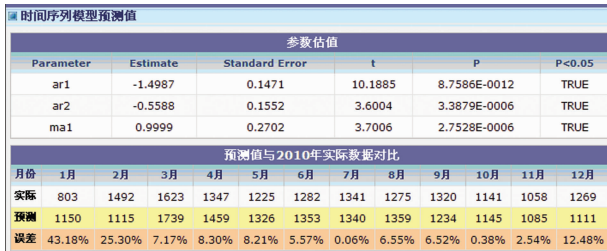


图 4 时间序列模型预测值图

2.3.3 调用灰色模型 (1)主要调用代码:.....conn.assign("gmData",gmData);//原始数据赋值 REXPInteger rexpInt=

new REXPInteger(startYear); conn. assign("startYear", rexpInt); //赋值起始年份整数值。……conn. voidEval(cmd); //执行 R 语言脚本命令。RList rlist=conn. eval("gmResult"). asList(); //取出预测值。(2)程序界面截图,见图 6~8。

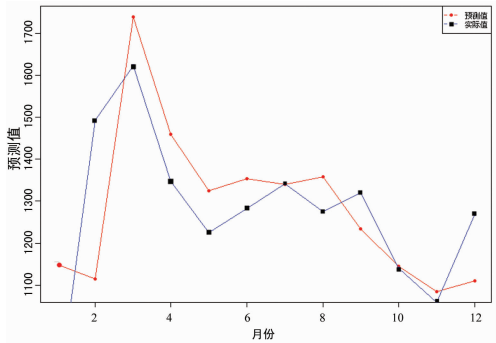


图 5 时间序列模型预测值与实际登记数对比折线图



图 6 灰色模型预测输入界面

登记病例数与预测值												
年份	1993年	1994年	1995年	1996年	1997年	1998年	1999年	2000年	2001年	2002年	2003年	2004年
实际值	8135	10329	11174	12264	15781	14387	15674	16402	18018	17187	23505	28836
基于2010年预测未来2年												
预测值												

图 7 灰色模型预测值数据表

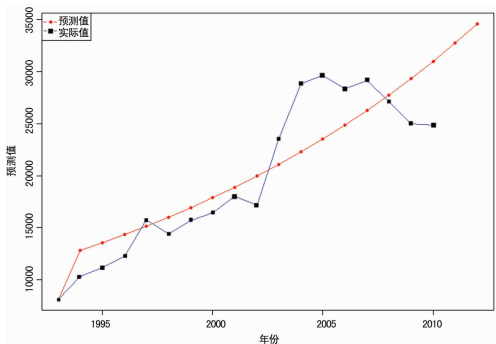


图 8 灰色模型预测值与实际登记数据对比折线图

结核病发病登记数据												
年份	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
数量	8135	10329	11174	12264	15781	14387	15674	16402	18018	17187	23505	28836
2009年和2009年的登记数据与预测值对比												
年份	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
登记病例数	25010	24868	-	-	-	-	-	-	-	-	-	-
Pmax	29644	29644	-	-	-	-	-	-	-	-	-	-
归一化预测结果	0.85259	0.83889	-	-	-	-	-	-	-	-	-	-
预测结果	25274	24868	-	-	-	-	-	-	-	-	-	-
误差	1.06%	0.00%	-	-	-	-	-	-	-	-	-	-
2011年预测结果												
2011	-	-	-	-	-	-	-	-	-	-	-	-

图 9 神经网络预测值

2.3.4 调用神经网络模型 (1)主要调用代码:……conn. assign("P",P); //绑定输入条件 P,1993 年至 2007 年发病数。conn. voidEval(cmd); //执行外部 R 脚本。predictValues = conn. eval("y"). asDoubles(); //以数组形式取出预测值。(2)程序界面截图,见图 9。由于神经网络预测,存在一定的随机性,故每次的预测值不一定相同。

2.3.5 数据管理 JDBC 为 Java 程序提供了简单的数据访问途径,可操作多种关系数据库^[10-12]。本文描述的数据管理以数据增、删、改、查为主,针对每位患者的基础数据,提供简易的统计功能。当对数据进行编辑时,某些特定属性(字段),如:性别、民族、职业等直接使用下拉列表选择完成,无需每次都输入复杂的中文;统计功能可根据登记日期、年龄范围、性别、职业、民族等作为统计所需的数据查询条件,可根据年份、月份、诊断分型、职业、年龄性别等进行不同条件的分组统计。

3 讨论

通过使用 Java 语言整合 R 语言分析预测服务,通过可视化的 WEB 交互界面,将原有晦涩难懂的 R 命令行操作界面转换为可视化的 WEB 交互界面,可完成使用不同的分析预测模型对相同的数据进行预测。大大降低了 R 语言的使用入门门槛,使大多数不具备计算机编程语言知识的人士能够轻松地对数据进行分析预测,用于指导实际工作的开展。

对于从事计算机专业工作人士,可参考本程序,将 R 语言无缝地集成到信息系统中,使其具备专业的分析预测功能。R 语言具备丰富的预测分析模型,截止 2011 年 3 月共有 4300 个,涉及经济、工业和卫生等各行各业,作为一款优秀的、免费开源的、功能强大的统计分析软件,但呈现给用户的都是晦涩难懂的命令行操作界面,缺乏简单易用的界面大大地降低了该软件的可用性。随着计算机硬件技术和软件工具的持续发展,R 也会在保持目前成就的情况下,通过不断改善计算环境,成为有力的科研工具^[3,13-14]。

下一步的研究方向是实现 R 语言的可视化操作:抽象出常用的分析预测模型调用方式,提供一个通用的分析预测交互界面,并根据各个分析预测模型的个性化参数,提供相应的选项输入界面;对于需要多个步骤才能完成的分析预测模型,采用向导式指引操作;提供数据源管理,以可视化的方式从文件或关系数据库中直接抽取数据进行筛选、类型转换或其他操作后进行分析预测;提供便捷的预测数据图表配置,根据不同行业的分析惯例,生成不同的数据表及相应的图形,如:区域图、饼图、条状图、线性图、K 线图、甘特图、时间序列图和表盘图等。

参考文献:

[1] 吴笃卿,胡东风,吴晔.中国疾病预防控制中心信息系统的结构与功能[J].地方病通报,2007,22(2):65-67.

[2] 石发恩,任如山,蒋达华.R 语言在消声器设计试验中的应用[J].金属矿山,2008(11):97-99,151.

[3] 罗玫,赵嵩正,蒋建洪.模糊综合评价模型的 R 语言实现[J].航空计算技术,2011,41(4):56-58,66.

[4] 张伟,史健勇,曹奇英.基于 R 语言健康成本测算分析系统的设计开发[J].上海工程技术大学学报,2011,25(3):276-280.

[5] 王俊宁,陈世铭,许金雷,等.基于 GPL 软件的小型生物实验室计算服务解决方案[J].生物信息学,2011,9(1):16-19.

[6] 杨树林,胡洁萍.Java EE 企业级架构开发技术与案例教程[M].北京:机械工业出版社,2010.

[7] 鲍振军,赵武,荣先奎.利用 JavaEE 构建创新型企业管理系统[J].机械设计与制造,2011(2):229-231.

[8] 邢丹丹,李劲华,李超.基于组合测试技术的 JavaEE 多层框架的测试方法[J].青岛大学学报:自然科学版,2011,24(3):38-41,64.

(下转第 1459 页)

检测细胞的增殖与凋亡,结果显示 5-Aza-CdR 可以明显抑制细胞增殖,诱导细胞周期阻滞和凋亡,细胞凋亡率明显提高。0.5 $\mu\text{mol/L}$ 的 5-Aza-CdR 对 T_{24} 细胞生长已有抑制作用,且随着药物浓度的增加,抑制作用逐渐增强,5.0 $\mu\text{mol/L}$ 5-Aza-CdR 作用 T_{24} 细胞 72 h 后,细胞的生长抑制率达到 23.0%。FCM 检测结果显示 5-Aza-CdR 处理的 T_{24} 细胞 G_0/G_1 期细胞明显增加,S 期细胞减少,凋亡率达到 $(40.35 \pm 3.63)\%$ 。细胞凋亡受众多凋亡相关基因调控,具有两条独立的途径:死亡受体途径和线粒体介导的凋亡途径,抑癌基因的表达状况是癌细胞凋亡的重要调节因素。本研究发现在人膀胱癌 T_{24} 细胞系中 Apaf-1 呈半甲基化状态,APC 基因的启动子呈高甲基化状态,而 5-Aza-CdR 作用后的 Apaf-1 及 APC 基因均呈去甲基化状态。基因启动子高甲基化或半甲基化可导致基因表达沉默或降低,是肿瘤发生的早期事件,而 DNA 甲基化是由 DNA 甲基转移酶催化的,5-Aza-CdR 抑制肿瘤细胞生长、诱导其凋亡的作用机制并非其本身的药物毒性直接所致,而是其去甲基化的间接作用^[10]。5-Aza-CdR 通过与 DNA 甲基转移酶共价结合,降低 DNA 甲基转移酶的生物活性,在 DNA 复制过程中可逆转基因 CpG 岛甲基化,调节基因表达,从而降低甲基化水平,因而被作为去甲基化制剂^[11-12]。已经证实 5-Aza-CdR 在多种肿瘤中具有抑制细胞增殖和促凋亡作用^[13-15]。本实验中发现,5-Aza-CdR 能成功逆转两基因启动子的甲基化状态,使因甲基化失活的 Apaf-1 及 APC 基因去甲基化而重新表达的同时,膀胱癌细胞出现增殖受限,凋亡明显增加,这进一步说明了 Apaf-1、APC 基因表达上调可能是 5-Aza-CdR 诱导 T_{24} 细胞凋亡和抑制细胞生长的重要机制之一。

通过本实验证实,5-Aza-CdR 可在体外诱导膀胱癌细胞凋亡和抑制细胞增殖,提示表观遗传学调控可能成为膀胱癌治疗新的靶点。恶性肿瘤中因基因突变、缺失等遗传学改变导致的基因失活难以逆转,而启动子甲基化所致的基因沉默是可逆的,可通过相应药物逆转表观遗传学的改变,恢复抑癌基因的功能,促进肿瘤细胞凋亡,达到治疗肿瘤的目的。目前,5-Aza-CdR 已经用于白血病、骨髓增生异常综合征的临床治疗。因此,如能将 5-Aza-CdR 与其他干预药物联合用药,可能为膀胱癌的治疗提供了新的思路。

参考文献:

[1] Esteller M. Epigenetics in cancer[J]. N Engl J Med,2008,358(11):1148-1159.
[2] Cabello MJ, Grau L, Franco N, et al. Multiplexed methylation profiles of tumor suppressor genes in bladder cancer[J]. J Mol Diagn,2011,13(1):29-40.
[3] Christoph F, Kempkensteffen C, Weikert S, et al. Frequent epigenetic inactivation of p53 target genes in seminomatous and nonseminomatous germ cell tumors[J]. Cancer

Lett,2007,247(1):137-142.
[4] Christoph F, Hinz S, Weikert S, et al. Comparative promoter methylation analysis of p53 target genes in urogenital cancers[J]. Urol Int,2008,80(4):398-404.
[5] Ahmad ST, Arjumand W, Seth A, et al. Methylation of the APAF-1 and DAPK-1 promoter region correlates with progression of renal cell carcinoma in North Indian population[J]. Tumour Biol,2012,33(2):395-402.
[6] Christoph F, Kempkensteffen C, Weikert S, et al. Methylation of tumour suppressor genes APAF-1 and DAPK-1 and in vitro effects of demethylating agents in bladder and kidney cancer[J]. Br J Cancer,2006,95(12):1701-1707.
[7] Iacopetta B, Grief F, Li W, et al. APC gene methylation is inversely correlated with features of the CpG island methylator phenotype in colorectal cancer[J]. Int J Cancer,2006,119(10):2272-2278.
[8] 王文卫,陈凌武,徐盈斌,等. 凋亡蛋白酶活化因子-1 基因及结肠腺瘤样息肉肉病易感基因启动子甲基化状态在膀胱癌组织中的表达[J]. 中华实验外科杂志,2009,26(2):229-231.
[9] 王文卫,陈俊星,潘俊,等. 曲古抑菌素 A 对膀胱癌 T_{24} 细胞增殖及 Apaf-1 和 APC 基因表达的影响[J]. 中山大学学报:医学科学版,2010,31(4):552-556.
[10] 崔军,于秀月,刘勇,等. 5-氮-2'-脱氧胞苷对 T_{24} 膀胱癌细胞 RUNX3 基因去甲基化的转录调节作用[J]. 山西医药杂志,2009,38(1):16-18.
[11] Kantarjian H, Oki Y, Garcia-Manero G, et al. Results of a randomized study of 3 schedules of low-dose decitabine in higher-risk myelodysplastic syndrome and chronic myelomonocytic leukemia[J]. Blood,2007,109(1):52-57.
[12] Chuang JC, Warner SL, Vollmer D, et al. S110, a 5-Aza-2'-deoxycytidine-containing dinucleotide, is an effective DNA methylation inhibitor in vivo and can reduce tumor growth[J]. Mol Cancer Ther,2010,9(5):1443-1450.
[13] 孟春风,彭过,戴冬秋,等. 5-氮杂-2'-脱氧胞苷诱导胃癌细胞系中 p16 和 MGMT 基因去甲基化并激活其表达[J]. 中国医科大学学报,2008,37(4):519-521.
[14] Issa JJ. DNA methylation as a therapeutic target in Cancer[J]. Clin Cancer Res,2007,13(6):1634-1637.
[15] Momparler RL. Epigenetic therapy of Cancer with 5-Aza-2'-deoxycytidine (decitabine) [J]. Semin Oncol,2005,32(5):443-451.

(收稿日期:2012-09-08 修回日期:2012-12-22)

(上接第 1456 页)

[9] Kabacoff RI. R in action[M]. New York: Manning Publications Co,2009.
[10] Richardson WC, Avondolio D, Schrager S, et al. Professional Java JDK[M]. 6 ed. New Jersey: John Wiley & Sons,2007.
[11] 张俐,张维玺. 改进的 JDBC 框架在数据持久层的应用[J]. 计算机工程与设计,2010,31(8):1746-1749.
[12] 李贤华,彭真. 基于 JDBC-ODBC 的数据库访问方法的设计

计与实现[J]. 计算机与现代化,2008(5):28-31.
[13] 胡海峰,鲍福廷,魏瀚,等. 固体火箭发动机试验数据分析[J]. 弹箭与制导学报,2011,31(1):131-134.
[14] 陈甫,余朋,李力. 成都双流机场历年主导能见度数据的统计分析 & R 实现[J]. 电脑知识与技术,2012,8(27):6428-6433.

(收稿日期:2012-09-18 修回日期:2013-01-26)