

# WSARE 计算方法在传染病暴发早期预警中的应用

张子武 冯子健 李晓松

**【导读】** 探讨异常模式探测方法——WSARE (What's Strange About Recent Events) 在传染病暴发早期预警中的应用价值, 拓展传染病病例监测数据的多维聚集性探测统计方法。分别采用基于历史数据和贝叶斯网络为基线的 WSARE 算法, 对 2007 年深圳市宝安区麻疹发病模拟实时监测预警。结果表明 WSARE 算法能够早期探测到麻疹在特定人群的异常增高, 在传染病暴发早期预警中具有重要应用价值。

**【关键词】** 传染病监测; 早期预警; WSARE 方法

**A study regarding the applicability of WSARE algorithms in the early warning system of infectious disease outbreaks** ZHANG Zi-wu<sup>1</sup>, FENG Zi-jian<sup>2</sup>, LI Xiao-song<sup>3</sup>. 1 Sichuan Center for Disease Control and Prevention, Chengdu 610041, China; 2 Chinese Center for Disease Control and Prevention; 3 Department of Health Statistics, Sichuan University  
Corresponding author: LI Xiao-song, Email: lixiaosong1101@126.com  
This work was supported by grants from the National Health Industry Research Project (No. 200802133) and National Natural Science Foundation of China (No. 30571618).

**【Introduction】** To investigate the application of WSARE (What's Strange About Recent Events) algorithm in early warning on outbreaks of infectious diseases and to explore the multi-dimensional statistical methods for the detection of infectious diseases outbreak. Using WSARE algorithms based on historical data and Bayesian Network as baseline respectively, to analyze data on measles by mimicking the real-time monitoring and early warning system in Bao'an district, Shenzhen city, in 2007. WSARE algorithms were considered to be effective and timely in detecting the abnormally increase of measles among special population. WSARE algorithm could timely detect the abnormal increase of diseases among special local populations, thus having important value in the application of early warning system during the outbreak of infectious diseases.

**【Key words】** Infectious disease surveillance; Early warning; WSARE algorithm

近年来在针对生物恐怖袭击、传染病暴发病状监测及早期预警系统领域<sup>[1,2]</sup>发展了许多先进的统计模型<sup>[3-13]</sup>,对探测传染病或症候群的聚集性具有重要应用价值,但也面临进一步解决算法优化和维度拓展等问题。近来 Wong 等<sup>[14-17]</sup>提出一种基于规则的异常模式探测方法(What's Strange About Recent Events, WSARE),用于疾病暴发的早期探测。该方法目前已纳入美国实时暴发监测系统(RODS system),其中基于历史数据基线的 WSARE 算法已

用于宾夕法尼亚州和犹他州急诊病例监测以及 2002 年盐湖城冬奥会的症状监测,基于贝叶斯网络基线的 WSARE 算法也被美国一些州的公共卫生部门以及以色列国家疾病预防控制中心使用<sup>[18-20]</sup>。WSARE 算法应用于非连续、多维的时序数据集,通过基于规则的方法将最近的数据与基线分布数据相比较,发现两者间比例变化最大的亚组,并通过随机化检验确定该差异产生的概率从而达到探测预警的目的。此外,通过贝叶斯网络将时间趋势纳入产生的基线分布,将季节效应、周末效应等时间因素也纳入分析。因此,WSARE 是一种融合了包括规则算法、贝叶斯网络、随机化检验等多种思想的聚集性探测算法。2004 年我国建立了传染病疫情信息网络直报系统,本研究旨在探讨 WSARE 算法对我国该系统数据的适用性,及对传染病暴发早期预警的应用价值。

DOI: 10.3760/cma.j.issn.0254-6450.2010.11.024

基金项目: 国家卫生行业科研专项经费(200802133); 国家自然科学基金(30571618)

作者单位: 610041 成都, 四川省疾病预防控制中心(张子武); 中国疾病预防控制中心(冯子健); 四川大学华西公共卫生学院卫生统计学教研室(李晓松)

通信作者, 李晓松, Email: lixiaosong1101@126.com

基本原理

WSARE算法基于历史数据和贝叶斯网络的基线,其区别主要在于基线分布的建立,前者以历史数据选定天数作为基线,后者是从构建的贝叶斯网络中抽样构造基线,在建立基线分布后,对最近发生的事件进行搜索,用基线和最近数据集寻找最佳得分规则,再用随机化检验估计最佳规则的P值,以得到有意义的P值及规则作为预警信号。

1. 基线构建: WSARE算法目的是探测最近事件中发生的异常。其中“最近事件”定义为所有纳入评价时间段的记录。该定义并非严格限定,而可以为任意时段。为了定义“异常”,首先需要确定“正常”,即基线的概念,在以历史数据为基线时,基线时间段的选取一方面要与“最近事件”足够近,以获取季节效应或临近趋势;另一方面又要与“最近事件”有足够的间隔,以避免最近发生了暴发而尚未检测到,因为如果基线时间段与“最近事件”距离太近,会随时间进展迅速将暴发病例合并入基线。例如为避免环境效应如周末效应,可选取一段时间之前星期数相同的几天数据作为基线,显然基线时间段需根据不同疾病而改变。

基于贝叶斯网络的基线则是运用贝叶斯网络产生基线的联合概率分布,该联合分布代表了包含环境属性的条件基线分布,环境属性即构成数据中各种趋势的属性,如季节趋势、周末效应等,需基于疾病的专业知识进行设定。基线分布学习包括: 纳入评价时间段以前的所有记录建立数据集,通过数据集建立贝叶斯网络。Moore等<sup>[21,22]</sup>发展了“基于ADTrees选择性重插法(optimal reinsertion based on ADTrees)”的算法,并运用于WSARE。该法首先从有向无环图中选择节点T,并断开各节点的连接,然后有效寻找最佳方法,根据得分函数将T重新插入有向无环图中,该法是一种大范围搜寻方法,有效的避免了局部最优。

完成贝叶斯网络学习可获得数据的联合概率分布,通过对环境属性值的限定,得到条件概率分布。然后从贝叶斯网络中限定环境属性值,抽取大样本的记录以代表条件概率分布,其中样本量要足够大以保证获得稀有属性组合的样本。

2. 最佳得分规则选取: 首先考虑单成分规则,取出第i天的相关记录,可得到某个属性的每个取值组合下的最近事件计数C<sub>recent</sub>和基线计数C<sub>baseline</sub>,构成一个四格表,即得到一个规则。以χ<sup>2</sup>检验比较最近事

件计数C<sub>recent</sub>和基线计数C<sub>baseline</sub>两组间分布的差异,与后面随机化检验所得P值区分,这里χ<sup>2</sup>检验产生的P值称为规则得分(因探测的是异常情况,通常采用Fisher精确概率法,且只关心某种情况数量的增加,故使用单侧检验)。

以BR<sup>1</sup>表示第i天单成分最佳得分规则,在BR<sup>1</sup>基础上寻找两成分最佳得分规则。将BR<sup>1</sup>与其他属性值配对,仍通过χ<sup>2</sup>检验得到规则得分,其中C<sub>baseline</sub>和C<sub>recent</sub>的计数为满足两成分规则的记录数,以BR<sup>2</sup>表示第i天的两成分最佳得分规则。设BR<sup>1</sup>的单成分属性值C<sub>1</sub>=V<sub>1</sub>,BR<sup>2</sup>的两成分为C<sub>1</sub>=V<sub>1</sub>,C<sub>2</sub>=V<sub>2</sub>在寻找两成分最佳得分规则时将两个成分的属性值进行组合,通过假设检验决定两个成分是否均有统计学意义。如表1所示,分别将C<sub>1</sub>=V<sub>1</sub>加入C<sub>2</sub>=V<sub>2</sub>及将C<sub>2</sub>=V<sub>2</sub>加入C<sub>1</sub>=V<sub>1</sub>。若两表得分均有意义,则认为第i天的最佳得分规则是BR<sup>2</sup>;若任意一个表的得分无意义,则认为第i天的最佳得分规则是BR<sup>1</sup>。

表1 两成分四格表

C <sub>recent</sub> 满足C <sub>1</sub> =V <sub>1</sub> 和C <sub>2</sub> =V <sub>2</sub> 的记录	C <sub>baseline</sub> 满足C <sub>1</sub> =V <sub>1</sub> 和C <sub>2</sub> =V <sub>2</sub> 的记录
C <sub>recent</sub> 满足C <sub>1</sub> ≠V <sub>1</sub> 和C <sub>2</sub> =V <sub>2</sub> 的记录	C <sub>baseline</sub> 满足C <sub>1</sub> ≠V <sub>1</sub> 和C <sub>2</sub> =V <sub>2</sub> 的记录
C <sub>recent</sub> 满足C <sub>1</sub> =V <sub>1</sub> 和C <sub>2</sub> ≠V <sub>2</sub> 的记录	C <sub>baseline</sub> 满足C <sub>1</sub> =V <sub>1</sub> 和C <sub>2</sub> ≠V <sub>2</sub> 的记录
C <sub>recent</sub> 满足C <sub>1</sub> ≠V <sub>1</sub> 和C <sub>2</sub> ≠V <sub>2</sub> 的记录	C <sub>baseline</sub> 满足C <sub>1</sub> ≠V <sub>1</sub> 和C <sub>2</sub> ≠V <sub>2</sub> 的记录

3. 规则P值的确定: 上述算法存在多重检验问题,需采用随机化检验解决。以UCP<sub>i</sub>表示原始数据未调整的P值,如前所述的规则得分。对迭代次数j=1~1000,以DB<sub>i</sub><sup>(j)</sup>表示新的随机化数据集,BR<sub>i</sub><sup>(j)</sup>表示DB<sub>i</sub><sup>(j)</sup>的最佳规则,UCP<sub>i</sub><sup>(j)</sup>表示BR<sub>i</sub><sup>(j)</sup>的未调整的P值(得分),以CPV<sub>i</sub>表示BR<sub>i</sub>的调整后P值。

$$CPV_i = \frac{\text{随机化检验中 } UCP_i^{(j)} < UCP_i \text{ 次数}}{\text{随机化检验次数}} \quad (1)$$

在决定是否预警时,采用的是调整后P值CPV<sub>i</sub>,而不是未调整P值(得分)UCP<sub>i</sub>。

由于随机化计算的计算量极其庞大,这里采用Maron和Moore<sup>[23]</sup>提出的优化算法以加快随机化检验的速度。如果在进行1000次迭代运算的早期能够确定CPV<sub>i</sub>>0.1,则可提前终止运算。如已进行了j次迭代运算,CPV<sub>i</sub>表示目前的CPV<sub>i</sub>值,具体算法同上。设CPV<sub>i</sub>为正态分布,可得CPV<sub>i</sub>的标准差σ<sub>CPV<sub>i</sub></sub>,并可通过

$$CPV_i \pm \frac{1.96\sigma_{CPV_i}}{\sqrt{n}} \quad (2)$$

计算CPV<sub>i</sub>的95%CI,如其下限

$$CPV_i - \frac{1.96\sigma_{CPV_i}}{\sqrt{n}} \quad (3)$$

>0.1,则可认为该概率在0.1水平上无意义。

对一般的数据集,暴发的概率较小,绝大多数天数中都将得到无意义的P值,因此通过优化算法可在这些天中提早终止迭代运算,加快随机化检验的速度。

### 实例分析

1. 数据来源及预处理:本研究以深圳市宝安区麻疹发病为例进行预警分析。传染病病例数据来源于我国疾病预防控制信息系统中深圳市宝安区2005—2007年麻疹确诊病例数据,经清理及逻辑查错,剔除重复卡、已删除卡及现住地址不详的数据,共计病例5741例。提取性别、出生日期、现详细住址、发病日期等信息,将数据转换成所需变量,即时间变量、空间变量、患者性别、年龄和职业。其中,以“发病时间”作为时间变量纳入分析,并得到相应的星期数及季节;以“现住址编码”所对应的共12个街道为空间变量纳入分析;将患者按年龄分为<1、1~5、6~11、12~17、≥18岁5个年龄段纳入分析;患者职业分为幼托儿童、散居儿童、学生、工人、家务及待业、其他。

2. 预警实施及参数设置:本研究分别采用基于历史数据和贝叶斯网络基线的WSARE算法,对2005—2007年深圳市宝安区麻疹数据进行分析,以2005—2006年麻疹数据作为历史基线数据,对2007年数据模拟逐日实时预警分析,即假设分析从2007年1月1日开始,每日纳入当天及当天之前的数据进行分析,向前逐日进行探测预警至12月31日。采用WSARE软件完成分析<sup>[24]</sup>。

首先采用基于历史数据基线的WSARE算法,由于麻疹的潜伏期一般为1~2周,因此基线时间间隔选为2周,基线时间长度为4周,考虑到由于数据

量较小,因此将4周内每天的数据均纳入基线。因探测的是异常情况,四格表中常有较小的数据,得分算法采用Fisher精确概率法。在基于贝叶斯网络基线的WSARE算法中最关键的是环境属性变量的设置,本研究所设置的是“day\_of\_week”(星期数)及“season”(季节)。由贝叶斯网络产生的基线样本量为10 000。参数设置见表2。

3. 预警结果:采用基于历史数据基线的WSARE算法模拟探测结果如表3。随机化检验 $\alpha=0.05$ 。探测到全年异常增高共13次,4、5、10、12月各1次,7、9、11月各2次,8月3次。其中发现双特征变量联合异常增高情况4次,单特征变量异常增高情况9次。采用基于贝叶斯网络基线的WSARE算法模拟探测结果如表4。探测到全年异常增高共46次,其中发现双特征变量联合异常增高情况24次,单特征变量异常增高情况22次。预警次数最多月份分别为12月8次,8、10月各6次。

模拟实时探测预警结果中,有现实意义较强的预警信号如基于贝叶斯网络基线的WSARE算法探测到2007年4月22—27日发病升高,通过回顾性分析发现该时段宝安区共发生麻疹47例,占4月总发病人数的33.8%,其中散居儿童22例(年龄1~5岁19例);地区分布表明,发病人数最高的松岗街道和西乡街道各发生16、12例。发病数远高于2005、2006年同期发病数,因此极有可能是一次麻疹的局部暴发,模拟探测预警也相应出现了连续报警。

一些预警信号需进一步确认分析,因为发出预警信号并不一定提示满足该特征变量的人群发病数量较高,而是由于其在随机化检验中显著高于基线的分布,例如2007年12月26日基于贝叶斯网络基线的WSARE算法探测到该市福永街道年龄<1岁的婴儿发病升高,当日该街道的婴儿实际发病人数

表2 两种算法的参数设置

算法	参数	含义	参数值
基于历史数据基线的WSARE	weeks	基线时间段的星期数	4
	weeks_offset	从多少星期前开始纳入基线	6
	days_in_week	基线时间段中每周纳入的单位天数	1
	scoretype	得分算法	FISHEREXACT_SCORETYPE
	num_randomizations	随机化检验迭代次数	100
	max_components	规则的最大成分数目	4
基于贝叶斯网络基线的WSARE	env_atnames	环境属性变量	day_of_week,season
	scoretype	得分算法	FISHEREXACT_SCORETYPE
	num_randomizations	随机化检验迭代次数	400
	max_parents	贝叶斯网络最大复节点数目	4
	num_bayesnet_rows	由贝叶斯网络产生的基线样本量	10 000
	max_components	规则的最大成分数目	2

表3 基于历史数据基线的WSARE算法实时探测预警结果

报警序号	日期	特征变量(1)	特征值	特征变量(2)	特征值	得分	P值
1	2007-04-27	地区	西乡街道			-0.006 42	0.020
2	2007-05-09	职业	学生			-0.000 50	0.001
3	2007-07-01	年龄	<1岁			-0.003 33	0.030
4	2007-07-25	年龄	1~5岁	性别	女性	-0.003 44	0.049
5	2007-08-01	年龄	<1岁			-0.000 68	0.010
6	2007-08-09	地区	西乡街道	年龄	>18岁	-0.003 48	0.020
7	2007-08-22	年龄	6~11岁			-0.000 26	<0.001
8	2007-09-21	地区	石岩街道			-0.008 71	0.010
9	2007-09-30	地区	石岩街道			-0.000 16	<0.001
10	2007-10-20	地区	石岩街道	性别	女性	-0.001 58	0.020
11	2007-11-04	地区	观澜街道			-0.009 61	0.030
12	2007-11-26	年龄	<1岁			-0.004 30	0.010
13	2007-12-15	年龄	1~5岁	性别	女性	-0.000 77	0.010

表4 基于贝叶斯网络基线的WSARE算法实时探测预警结果

报警序号	日期	特征变量(1)	特征值	特征变量(2)	特征值	得分	P值
1	2007-01-01	地区	西乡街道			-0.000 14	0.005
2	2007-01-22	地区	新安街道			-0.000 60	0.030
3	2007-01-28	地区	西乡街道	性别	女性	-2.96E-5	<0.001
4	2007-04-22	地区	松岗街道	性别	女性	-0.000 12	0.002
5	2007-04-23	地区	松岗街道	年龄	6~11岁	-0.001 48	0.018
...	...	...	...	...	...	...	...
42	2007-12-10	地区	观澜街道	年龄	<1岁	-0.000 52	0.030
43	2007-12-22	地区	公明街道	年龄	<1岁	-6.81E-5	0.002
44	2007-12-23	地区	观澜街道			-4.56E-5	0.005
45	2007-12-24	地区	观澜街道			-1.46E-6	<0.001
46	2007-12-26	地区	福永街道	年龄	<1岁	-0.001 76	0.030

仅为2例,但全年该街道婴儿单日发病数最高即为2例,且仅出现2次,而作为基线的2005、2006年该街道全部12月龄婴儿发病数相加也仅为3例,因此虽不能肯定2007年12月26日在福永街道婴儿中发生了麻疹暴发,但当天该街道的确发生了发病“异常”。

两种方法均发出预警信号的天数共6 d,其中有5次预警信号的特征变量值相同。可见对于异常增加较明显的情况,即便基线选取的方法不同,依然能够将其发现,同时也说明这类预警信号通常具有较强的现实意义,例如2007年9月30日探测到该市石岩街道有麻疹异常增高,通过分析发现当日石岩街道共发病11例,为全年该街道单日发病人数最多一天,且全年单日发病数大于8例的仅此一天。

## 讨 论

1. 预警信号及其应对: WSARE算法比较的是“最近事件”(探测时段)与基线分布的区别是否有显著差异,即探测发病的异常增高,但发出预警信号并

不等同于流行病学意义上的“暴发”,因此对WSARE探测所得结果,需进一步结合现场调查及实验室检查对发病的流行情况进行评估。

本研究结果显示,探测预警模型发出的预警信号的现实意义根据具体情况有所不同,需要通过进一步分析探讨其价值并采取不同的应对措施。可尝试引入预警信号分级的方法,对现实意义不同的预警信号赋予不同的强度级别,采取相应的预案。当探测预警模型发出预警信号,首先通过对当日的原始数据分析对预警信号进行评估,对于预警信号现实意义较强的情况提出高级别警告,如本研究中疾病发病数量激增或显著高于历史同期水平,两种方法均发出预警信号,连续3日以上发出预警信号等。对于高级别警告应立即进行核实,排除误报等偶然情况,并立即组织流行病学现场调查和实验室检测,对于真实的传染病暴发,应立即启动传染病应急预案,采取适当的措施进行干预。一些预警信号具有一定现实意义,但尚不明确,首先通过当日发病数据分析进行初步判断,对于发病异常升高较明确的,应采取流行病学现场调查等相应措施,对于现实意义较弱的预警信息可提出低级别警告,并予以关注。对于无现实意义的预警信号应予以排除。

2. 变量设置与参数设置: WSARE算法只能用于分类变量数据,对于计量变量数据,可在数据预处理时将其转换为分类数据,但这种转换将损失部分信息,特别是分类的截断值选取对结果有较大影响。本研究变量的转换主要依据专业知识,并可通过比较产生的不同结果来选取截断值。

在变量的选择上,可根据不同的研究需要、病种和资料可及性进行选择,对于贝叶斯网络基线的算法除了作为探测因素的反应变量,还包括影响反应变量变异的环境变量。环境变量的选取范围较广,除了本研究中纳入的星期数、季节,还可包括气候因素、食品用水条件、药物可及性等。但需注意,随着反应变量的增加,基线所需数据量和时间也相应增加,对于贝叶斯网络基线的算法,环境变量越多,贝叶斯网络结构越复杂,所需的基线时间也越长。

关于参数设置,对于历史数据基线的算法,关键的参数是基线时间段和时段内纳入的天数。如前所述,基线时间段的选取一方面要与“最近事件”足够相近,同时又要与“最近事件”有足够的间隔,这需要结合具体病种的传播特征等因素加以考虑。此外,选取的时间段内纳入基线的天数也会影响结果,如只纳入与探测当天相同星期数的数据,则可能避

免周末效应,但基线数量较少,如纳入选取的时间段内每日的数据,则会获得更多的基线信息,但不能控制周末效应。因此,对于基线时间段和时间段内纳入人数应根据不同疾病和数据特征进行选取。对于贝叶斯网络基线的算法,关键的参数是环境变量和贝叶斯网络学习参数,本研究中设定的环境属性变量为星期数和季节,贝叶斯网络最大复节点数目采用默认的 4,由贝叶斯网络产生样本数量为 10 000,取得较好的探测结果。同样对于该法参数设置需根据研究目的、数据条件及不同病种进行相应调整。

3. WSARE 算法的应用:Wong 等<sup>[25]</sup>通过模拟数据研究显示,WSARE 算法在多元分类资料异常情况探测中比单因素方法更为有效,其优势在于可识别不同特征变量组合构成的异常组群,而不是需要首先指定某些特征的组合进行监测,时效性和精确性优于控制图法及移动平均法。本研究通过实施模拟实时监测预警显示,WSARE 算法能够充分利用和挖掘多维度的传染病疫情数据,探测到局部地区特定属性人群的发病异常。一般研究仅是回顾性的对集合数据进行分析,无法针对发病个体监测信息进行实时预警,也无法发现不同维度属性组合人群的发病异常。研究结果显示,WSARE 算法在探测的时效性和精确性方面具有明显优势,对基于传染病网络直报系统数据的监测和预警具有应用前景。

不同基线的 WSARE 算法因方法和数据要求上的差异,在传染病监测预警中的应用有所不同,对于短期的监测预警,如大型会议和体育活动等,由于对数据收集的时间不长,而且一般不易获得环境属性资料,因此可采用基于历史数据基线的 WSARE 算法。而基于贝叶斯网络基线的算法是以长期的历史数据构造贝叶斯网络,并纳入环境属性,考虑了季节等时间趋势因素,对于长期监测预警,如对法定传染病、慢性疾病监测等,采用该算法具有明显优势。本研究显示,对于实时监测预警可结合两种算法的结果,以综合考虑长期和短期的波动异常。

#### 参 考 文 献

- [1] Stoto MA, Schonlau M, Mariano LT. Syndromic surveillance: Is it worth the effort. *Chance*, 2004, 17(1):19-24.
- [2] Reingold A. If syndromic surveillance is the answer, what is the question? *Biosecure Bioterror*, 2003, 1(2):77-81.
- [3] Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. *J R Stat Soc A Stat Soc*, 2001, 164(1):61-62.
- [4] Kulldorff M. Syndromic surveillance without denominator data: the space-time permutation scan statistic for disease outbreak detection. *PLoS Med*, 2005, 2:216-224.
- [5] Neill DB, Moore AW. Detecting space-time clusters: prior work and new directions. Technical Report, Carnegie Mellon University, 2004.
- [6] Neill DB, Moore AW. A fast grid-based scan statistic for detection of significant spatial disease clusters. *MMWR*, 2004, 53 (Suppl Syndromic Surveillance):255.
- [7] Neill DB, Moore AW. A fast multi-resolution method for detection of significant spatial disease clusters. *Advances in Neural Information Processing Systems* 16, MIT Press, 2004.
- [8] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, 1989, 77(2):257-286.
- [9] Strat YL, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. *Stat Med*, 1999, 18(24):3463-3478.
- [10] Rath TM, Sebastiani CM. Automated detection of influenza epidemics with hidden Markov models. *Advances in Intelligent Data Analysis*, 2003:521-531.
- [11] Cooper B, Lipsitch M. The analysis of hospital infection data using hidden Markov models. *Biostatistics*, 2004, 5(2):223-237.
- [12] Cooper G, Dash D, Levander J, et al. Bayesian biosurveillance of disease outbreaks. *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*. Banff, Canada: AUAI Press, 2004:94-103.
- [13] Wong WK, Cooper GF, Dash D, et al. Bayesian biosurveillance using multiple data streams. *MMWR*, 2005, 54(3):63-69.
- [14] Wong WK, Moore AW, Cooper GF, et al. Rule-based anomaly pattern detection for detecting disease outbreaks. 18<sup>th</sup> National Conference on Artificial Intelligence, 2002.
- [15] Wong WK, Moore AW, Cooper GF, et al. Bayesian network anomaly pattern detection for disease outbreaks. 20<sup>th</sup> International Conference on Machine Learning, 2003.
- [16] Wong WK, Moore AW, Cooper GF, et al. WSARE: What's strange about recent events? *J Urban Health*, 2003, 80(2):i66-75.
- [17] Cooper G, Dash D, Levander J, et al. Bayesian biosurveillance of disease outbreaks. *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence*. Banff, Canada: AUAI Press, 2004:94-103.
- [18] <http://www.autonlab.org/autonweb/16620.html>.
- [19] Gesteland H, Gardner M, Tsui FCh, et al. Automated syndromic surveillance for the 2002 Winter Olympics. *J Am Med Informat Ass*, 2003, 10(6):547-554.
- [20] Kaufman Z, Wong WK, Tamar PL. Evaluation of a syndromic surveillance system using the WSARE algorithm for early detection of an unusual, localized summer outbreak of influenza B: implications for bioterrorism surveillance. *IMAJ*, 2007, 9:3-7.
- [21] Moore A, Lee MS. Cached sufficient statistics for efficient machine learning with large datasets. *J Art Int Res*, 1998, 8:67-91.
- [22] Moore A, Wong WK. Optimal reinsertion: a new search operator for accelerated and more accurate Bayesian network structure learning. *Proceedings of the 20<sup>th</sup> International Conference on Machine Learning*, 2003:552-559.
- [23] Maron O, Moore AW. The racing algorithm: model selection for lazy learners. *Art Int Rev*, 1997, 11(1):193-225.
- [24] <http://www.autonlab.org>.
- [25] Wong WK, Moore A, Cooper GF, et al. What's strange about recent events (WSARE): an algorithm for the early detection of disease outbreaks. *J Mach Learn Res*, 2005, 6:1961-1998.

(收稿日期:2010-05-28)

(本文编辑:张林东)