利用遗传算法获取鸭类传染病环境因素的相关系数

王卫兵[1],杨崇俊[1],蒋之犇[2],王 勇[2],朱红缘[3],雷少华[4],徐 冰[2]*

(1. 中国科学院遥感应用研究,北京 100101;2. 清华大学,北京 100084;3. 首都师范大学,北京 100048;4. 犹他大学,美国犹他州 84112)

摘要　以调查的鸭类数据为例,基于统计方法得到养殖点周围 100 m 的土地利用类型:水田、旱地和水体以及养殖点到国道的距离、养殖点到铁路的距离、养殖点到湖泊的距离和养殖点到河流的距离等因素与鸭传染病的发生具有相关性。最后用遗传算法获得这 7 个环境因素的权重,得到鄱阳湖区养殖点的鸭传染病的发生与养殖点离国道的距离,养殖点离河流的距离等有最大的相关性。

关键词　遗传算法;遥感;地理信息系统

中图分类号　S858.32　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)28 - 13618 - 03

Correlation Coefficient of Environmental Factors Related to Infectious Diseases of Ducks by Genetic Algorithm

WANG Wei-bing et al　(Institute of Remote Sensing Applications (IRSA), Chinese Academy of Sciences, Beijing 100101)

Abstract　Taking the ducks data investigated as an example, based on statistical methods, different land use types around the culture points within 100 meters were obtained: paddy fields, dry fields and waters, as well as distance from culture points to the national road, distance from culture points to the railway, distance from culture points to the lakes and distance from culture points to the rivers have the relationship with the occurrence of ducks diseases. At last, Genetic Algorithm was used to obtain that weights of the seven environmental factors, as well as occurrence of infectious diseases of the ducks in Poyang Lake has the greatest relevance with distance from culture points to the rivers and distance from culture points to the national road.

Key words　Genetic Algorithm; Remote sensing; GIS

禽传染病的发生受到多种因素的影响,包括自然环境因素、社会环境因素、销售渠道因素[1-2]。土地利用类型对鸭类养殖点的分布和鸭类传染病的发生有影响。鸭类养殖地点一般在水塘、河流、湖泊边,离农田比较近,有时在稻田中放养,早、中、晚三季稻跟家鸭养殖有关系[3]。水稻收割后,放养的家禽数量增加,会提高家禽与外部环境接触的机会,提高禽流感发生的可能性。交通网络可能会影响鸭传染病的传播,如果与交通网络的联系紧密,那么鸭类与流动的人口和车辆废弃物可能有更多的接触机会,可以导致得病的几率变大。Trapman Pieter 等以控制交通作为一个重要的措施建立了家禽传染病控制模型[4]。不同特征的水系决定着候鸟的分布,而候鸟是禽流感以及其他传染病病毒的一个巨大病原体蓄积库和传播媒介,候鸟的分布与活动可影响鸭传染病的传播。带毒候鸟在迁徙沿途通过排泄物污染水源及土壤,可能造成鸭传染病的发生与传播。湿地和自然保护区是野生鸟类的主要繁殖地、越冬地和迁徙路线上的停歇地[5]。湿地和保护区周围的场所是候鸟觅食和活动的区域,如果家禽养殖在附近区域,与带毒候鸟接触后,鸭类的得病几率变大。养殖地区如果有很多人类流动,可能使鸭传染病病毒的传播更频繁,使家禽感染鸭传染病病毒的机会变大。家禽养殖者之间的交流也可能会导致病毒之间的传播。活禽的运输可使禽类跨地区流动,而活禽交易市场和农贸市场由于卫生条件普遍较差,可能会使病毒传播的更快。如 1997 年在香港特别行政区,拥挤的条件下的居民区附近活家禽市场迅速传播禽流感病毒[6]。

国内外研究者在对禽传染病进行研究的过程中,针对鸭类传染病如禽流感、鸭瘟、鸭霍乱等流行病学研究,意识到鸭类在其他鸟类传染病发生与传播中可能扮演重要角色。在亚洲,户外的家禽可能已经把禽传染病的病毒传染给野生鸟类[1]。鸭类可能通过直接接触受感染水禽或其他受感染家禽,或通过接触已被病毒污染的表面(如泥土或笼)或材料(如水或饲料)感染禽传染病的病毒。人、车辆和其他无生命的物体等可作为媒介,从一个农场到另一个农场传播禽传染病[1]。因此,笔者以鸭类为例,研究禽传染病的发生与环境因素的关系以及与禽传染病的发生有关的各种环境因素相关系数。

1　数据与方法

1.1　数据准备与处理　该试验的区域是鄱阳湖地区,笔者将研究区域进行划分,用 30 m 分辨率的 TM 影像的栅格来做。根据鸭传染病病毒的流行病学特性,可以提出众多的风险因素。鉴于研究目的,数据是否可以获取等方面的考虑,整理了如下的一些数据资料:养殖点周围 100 m 方形区域的各种土地利用类型百分比数据包括耕地、林地、草地、水体、居民区、未利用土地等,养殖点到国道、省道、高速公路、铁路、湖泊、河流、居民点距离共 13 项环境因素数据。

对于养殖点周边的土地利用类型百分比,利用邻域分析法得到更好的数值。距离值是要提取的一类重要指标,主要通过距离来研究鸭传染病的发生与水体、交通、居民点等因素的相关性。对于养殖点到上述各要素之间的距离,用(1)式的欧几里德距离来求解。

欧几里德距离定义为:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

式中,d 为平面上两点要素之间的距离。这里,(x_1, y_1) 和 (x_2, y_2) 分别是 2 个要素的平面位置坐标。

土地利用数据采用中国科学院遥感应用研究所根据 TM 影像数据得到的土地利用类型的数据;鸭类疾病数据:2008 年的数据是笔者调查得到的,2002 ~ 2007 年的数据是调查问卷的追溯数据;地理数据:用到了中科院遥感所获取的交通、

基金项目　国家高技术研究发展计划 - 863 计划(2009AA12Z227)。

作者简介　王卫兵(1983 -),男,山东东营人,硕士研究生,研究方向:网络空间信息系统,流行病学。* 通讯作者,E-mail:bingxu@tsinghua.edu.cn。

收稿日期　2009-05-15

水域、居民点等数据。对上述数据进行处理,得到养殖点周围 100 m 方形区域内各种土地利用类型的百分比以及养殖点到各种距离因素的最近距离数据。

1.2　统计计算　有病的养殖点和没有病的养殖点环境因素的距离近,而两者离省道、高速、居民点的距离差不多。因为养殖点距离交通线路近的得病多,可能交通运输起到一定作用。同样地,离湖泊和河流近的得病多,也可能与候鸟离开栖息地觅食的远近有关,离开河流和湖泊越远,候鸟到达的机会越小,得病几率越小。

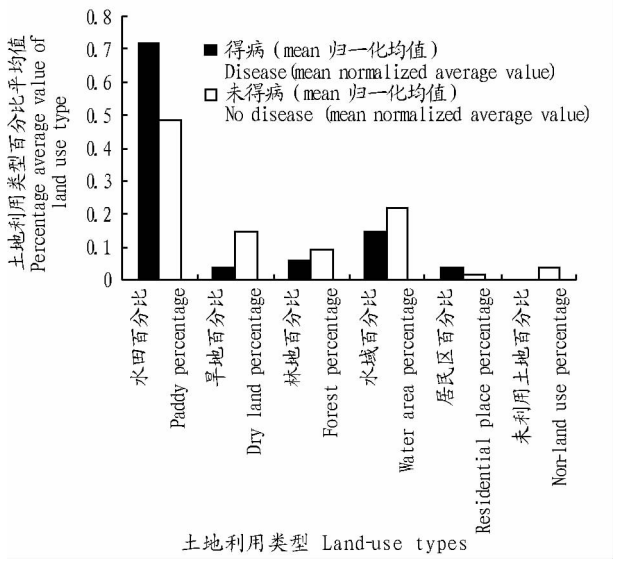

图 1 土地利用类型数据平均值的对比

Fig. 1　Contrast map of data average value of land use types

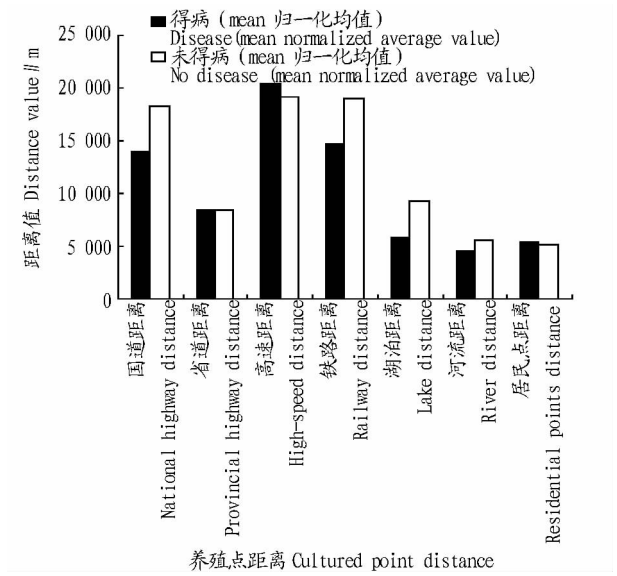

图 2　距离数据平均值的对比

Fig. 2　Contrast map of distance data average value

数据的平均值的对比见图 1、2。由图 1、2 可知,得病点的土地利用类型中水田的百分比所占的比重最大,水域的百分比所占的次之,旱地所占的很少。没得病点的土地利用类型中,水田的百分比所占的比重最大,水域的百分比所占的次之,旱地所占的很少。由于调查的养殖点的居民点、林地、草地、未利用土地的百分比绝大多数没有数据,该试验可以不考虑这些因素。通过对比得病点和没得病点的环境因子数据,得出得病的养殖点周围 100 m 的水田的百分比比没得病的大,旱地的百分比没得病的少,水域的百分比比没得病的少一点。其中,得病点的水田百分比比没得病的大,可能与野鸟来水田觅食容易接触鸭类有关。得病的养殖点距离国道、铁路、湖泊、河流的距离比没有得病的养殖点离这些因素

为了直观得到得病与没得病的环境因素差距,利用图 3 来表示两者之间的归一化数值差。从图 3 中看出,环境因素 1、2、3、4、5、6、7、10、11、12 的两者之差值具有明显的分类区分的意义。而 3、5、6 因为绝大多数调查的养殖点周围 100 m 内没有该数据,因而可以舍弃。最终得到环境因素 1、2、4、7、10、11、12 对养殖点得病与否具有决定作用,分别是水田的百分比,旱地的百分比,水域的百分比,养殖点离国道的距离,离铁路的距离值,离湖泊的距离,离河流的距离等因素。

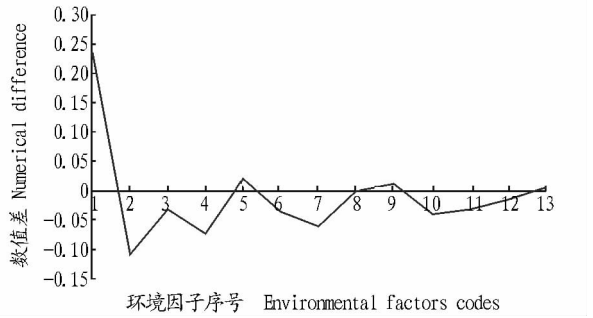

图 3　得病养殖点与没得病养殖点的环境因子的归一化数值差

Fig. 3　Normalized numerical difference of disease breeding points and no disease at the environmental factors

1.3　遗传算法求环境因子权重　遗传算法(Genetic Algorithm,GA)是其中一种非常有效的随机搜索方法,是一种基于遗传学机理的概率搜索技术。GA 一般通过初始化产生问题解的初始群体,然后用事先定义的适应度函数评价群体中的每个个体;以后每一代的个体都按照适应度函数的值进行选择,并且通过遗传算子的进化产生新的适应性更好的群体。通过这种机制遗传算法可以收敛到全局最优解或次优解[9]。该文研究基于 GA 来获取与鸭类传染病发生有关的环境因子的权重,得到鸭类传染病与环境因子的相关性。对于现在的养殖点数据,怎样利用它们求解环境因素的权重是一个问题。齐平曾经采用结合遗传算法与模拟退火算法的遗传模拟退火算法对特征项赋予权重[8]。其中,对于特征项权重的获取,引入了基于范例的搜索方法[7]。试验中,范例是养殖点,特征项是与传染病发生相关的各个环境因素,这样通过遗传算法就可以得到这些因素的权重。笔者统计得到了 7 个环境因素对养殖点传染病发生有大的作用,可以用这些环境因素对鸭传染病的发生进行预测。

相似性是衡量对象之间相似度的指标,一般通过计算对象在特征空间中的距离获得。在基于范例的推理系统中,大多数的范例检索都使用最近邻算法(*K-NN*)。试验中,*K-NN* 算法使用带权重的距离度量算法,即给样本数据集的每个属性赋以一定的权重,权重大小表示属性具有高低不同的相关性[9],利用带有权重的最近邻距离算法来计算适应度函数。算法在开始阶段,将需要确定属性的权值的数据库数据分成 3 部分:参考样本(*REF*)、测试样本 1(*TEST*1)和测试样本 2(*TEST*2)。$ref[i] \in REF$,$test1[i] \in TEST1$,$test2[i] \in TEST2$,

其中,$ref[i]$,$test[i]$,$test2[i]$分别表示参考样本和测试样本 1 和测试样本 2 中的第 i 个养殖点数据,参考样本中的养殖点总数为 m,测试样本 1 中的养殖点总数为 n,测试样本 2 中的养殖点总数为 s。笔者主要采用遗传算法的改进方法,防止解过早地收敛。在初始种群中,对所有的个体按其适应度大小进行排序。改进选择方式,放弃轮赌选择,以避免早期的高适应度个体迅速占据种群和后期种群中因个体的适应度相差不大而导致种群停止进化。采用相对最优选择法选取父个体,产生下一代,具体是从 20 个原始父个体中随机选 5 个,找出其中适应度最大的那个作为一个父个体,再从 20 个中任意选 5 个父个体,寻找适应度最大的那个作为另一个父个体,这 2 个交叉变异出 2 个子个体。以相同的方式产生 15 个子个体。

主要公式有:

$$D(X_i,Y_i)=|X_i-Y_i| \tag{2}$$

式中,$D(X_i,Y_i)$是 2 个养殖点(x 和 y)7 种对应的环境因素各自($X_i,Y_i,i=1,\cdots,7$)的归一化的数值之间的差。

$$DIS(X,Y)=[\sum_i W_i * D(X_i-Y_i)^r]^{1/r} \tag{3}$$

式中,假设样本点 $X=\{X_1,K,X_n\}$,$X_i(1\leqslant i\leqslant n)$ 是它的特征值,W_i 是其权重。X 是 n 维特征空间 $D=\{D_1,K,D_n\}$ 上的一点,$X_i\in D_i,n=7$。对于 D 上的 X,Y,则 X,Y 在 D 上的距离为:$DIS(X,Y)$。当式(3)中 $r=2$ 时,$DIS(X,Y)$为欧拉距离。试验中,$DIS(X,Y)$是 2 个养殖点(X 和 Y)7 种对应的环境因子($X_i,Y_i,i=1,\cdots,7$)的归一化的数值之间的差与相应的权重乘积的和。

参考样本的养殖点 k 和测试样本 1 中的养殖点 j 之间的加权距离:

$$DIS1[test(j),ref(k)]=\sqrt{\sum_f W[i]f\cdot D(test1[j]_f,ref[k]_f)^2} \tag{4}$$

参考样本的养殖点 k 和测试样本 2 中的养殖点 j 之间的加权距离:

$$DIS2[test(j),ref(k)]=\sqrt{\sum_f W[i]f\cdot D(test2[j]_f,ref[k]_f)^2} \tag{5}$$

试样本 1 中所有养殖点在所有的权重解中的第 i 个权重解情况下与参考样本中所有的养殖点之间总的加权距离的和:

$$T1[i]=\sum_{j=0}^{n}\sum_{k=0}^{m}DIS1(test1(j),ref(k)) \tag{6}$$

试样本 2 中所有养殖点在所有的权重解中的第 i 个权重解情况下与参考样本中所有的养殖点之间总的加权距离的和:

$$T2[i]=\sum_{j=0}^{n}\sum_{k=0}^{m}DIS2(test2(j),ref(k)) \tag{7}$$

2 遗传算法的算法过程

创建一个随机的初始状态,初始化 20 组 7 种环境因素的权重解。评估适应度,利用 K-NN 算法,找出测试样本 1 中每一个养殖点距离参考样本中最近的养殖点,判断得病信息是否相同,得到在每组权重解下,测试样本 1 的判断正确率,以此作为适应度函数的值。根据适应度值,按从小到大的顺序对 20 个权重解进行排序。

繁殖(包括选择,交叉和变异)。试验中,选择 5 个适应度值高的权重解直接复制到下一代。将 20 个权重解的各个个体随机搭配成对,对每一对权重解,以某个概率交换它们之间的部分数据;对 20 个权重解中的每一个,以某一概率改变某一个或某一些基因组上的基因值。试验中,取变异率为 0.1,剩余的 15 个子代是 20 个父代交叉和变异得来的。如果新一代包含一个解,能产生一个充分接近或等于期望答案的输出,那么问题就已经解决了。如果情况并非如此,新的一代将重复他们父母所进行的繁衍过程,一代一代演化下去,直到达到期望的解为止。停止循环。直到满足停止循环的条件。

3 试验具体过程

(1)读取数据文件,随机初始化 *REF* 含有 20 个养殖点,*TEST*1 含有 17 个养殖点和 *TEST*2 含有剩余的 17 个养殖点,即随机产生 ref[20],test1[17],test2[17]。

(2)随机产生权重解数组,即 weight[20];初始化 20 个权重解数组。

(3)对 weight[20]数组中的每个对象 weight[i]根据(3)式和(4)式计算参考样本和测试样本 1 之间的加权距离,再判定测试样本 1 中所有的养殖点的得病数据的正确率(即看测试样本 1 中养殖点数据项的最后一项得病与否的信息是否与参考样本中加权距离最近的养殖点得病信息一致,如果一致,则说明判断正确),令正确率的值作为适应度函数 fitter[i]。

(4)根据适应度值的大小,按从小到大的顺序对相应的权重值 weight[i]进行排序;并把最好的判断得病与否的正确率和对应的权重值记录下来。

(5)利用遗传算法对 weight[20]优化 100 代,选出最好的解。选择的过程是每迭代一次找出一个判断正确率最高的解,以后只要不是超过这个解的判断的正确率,都用此解作为最好的解。如果有超过此解正确率的解出现,用超过的解作为最好的解。

表 1 训练样本的分类结果

Table 1 Classification results of trained samples

实际的分类 Actural classification	预测的分类(Yes) Predicated classification (Yes)	预测的分类(No) Predicated classification (No)
Yes	0.6	0.4
No	0.1	0.9

表 2 7 个环境因子相应的权重值

Table 2 Corresponding weights of 7 environmental factors

水田百分比 Paddy percentage	0.004 2
旱地百分比 Dry land percentage	0.006 3
水体百分比 Water body percentage	0.001 0
国道 National highway	0.262 5
铁路 Railway	0.124 2
湖泊 Lake	0.069 2
河流 River	0.272 6

(6)运行一万次,得到一万次每次的最优解。再选出这一万个解中选择判断正确率最高的一些解,对这些解求训练样本 2 的判断正确率。如果有多个权重解判断正确率相同,

(下转第 13640 页)

剂量用量。一般临床推荐用量可参照经口给药量比例换算[5],具体换算比例如表2所示。

表2 人及不同动物的一般药物与植物凝集素临床推荐用量

Table 2 Clinical recommended dosage of general drug and PHA for human and different animals mg/kg

物种 Species	一般药物用药量 General drug dosage	植物凝集素(安全剂量) PHA (safe dosage)
人 Human	1	1.25
小鼠,大鼠 Small mouse, big mouse	50 ~ 100	125 ~ 129
豚鼠 Guinea pig	15 ~ 20	18 ~ 25
猫,狗 Cat, dog	5 ~ 10	6 ~ 12

在非肠道途径给药时,上述换算比例应适当减小[5-6]。

麦佩瑜等[7]1994年对广州市医药工业研究所生产的冻干粉针剂PHA进行了急性毒性试验,结果是PHA LD_{50} = 1 356.4,95%可信区间1 198 ~ 1 536 mg/kg,这与该试验结果有很大差异,分析原因认为可能是如下条件因素所致:①所用原料不同生物活性不同,从云南丽江白芸豆中提取PHA,活性测定表明其凝聚活性为3.0 μg/ml;麦佩瑜等文章中使用批号为870423,活性没有报道,他们试验时用的PHA时间放置过长(因为相关网站报道的保存时间是2a)[8]。②由于不同种类,不同来源的植物凝集素其作用不同,会引起急性毒性试验结果不相同。③提取制备工艺不同,会影响植物凝集素的效价,也会导致结果差异。

目前在国内植物凝集素临床用于急性白血病,恶性肿瘤,病毒性肝炎,乙型脑炎等等。给药的剂量和途径是成人每日20 ~ 40 mg,儿童稍减,以生理盐水或葡萄糖液稀释后静脉注射[9]。该试验研究得到的结果与王海燕等[9]基本吻合。

3 小结

试验结果表明,自行批量提取的云南丽江大白芸豆植物凝集素对昆明小白鼠的 LD_{50} 是140.24 mg/kg,95%置信区间为125.19 ~ 178.91 mg/kg。属于中等到低等性,使用时应遵守安全剂量范围。该植物在云南丽江产量大,品质优,所提取的植物凝集素效价高,有广阔的开发前途。

参考文献

[1] 侯建军,肖玄,黄邦钦. 细胞凝集素的生物学特性及应用研究进展[J]. 湖北民族学院学报:自然科学版,2004,22(3):64

[2] 舒晓燕,阮期平,侯大斌. 植物血凝素的研究应用[J]. 现代中药研究与实践, 2006,20(6):53 - 56.

[3] 吴叶,王昌梅,张丽芬. 人及兔醛化红细胞的制备和应用[J]. 安徽农业科学,2008,36(21):8885 - 8888.

[4] 楼宜嘉. 药物毒理学[M]. 2版. 北京:人民卫生出版社,2007,176 - 178.

[5] 刘毓谷. 卫生毒理学基础[M]. 北京:人民卫生出版社,1987:87

[6] 工业毒理学实验方法编写组. 工业毒理学实验方法[M]. 上海:上海科学技术出版社,1979:6,341.

[7] 麦佩瑜,廖仕元,张维政. 毒性实验研究[J]. 广东医药学院学报,1994,10(4):258 - 259.

[8] http://www.gzpiri.com/website/5/2005621154024.htm .

[9] 王海燕,白晓春,罗深秋. 植物凝集素与医学应用[J]. 生命的化学,2003,23(3):224 - 225.

(上接第13620页)

再采用(5)式和(7)式求相似性最近的解,此权重解为运行一万次之后的最优解。

4 结论与讨论

(1)经过遗传算法的试验,最好的权重解结果为表1,它总的预测正确率为0.764 7。其中判断0(没得病)的正确率和判断1(得病)的正确率以及判断0和判断1的错误率如表2。结果显示,7个环境因素中,养殖点离国道的距离,离河流的距离的权重最大。

(2)经过统计方法,得到候鸟的活动区域的土地利用类型(水域、耕地类型等),养殖点到交通线路(国道、铁路等),河流等的距离与鸭传染病发生具有相关性。根据这组数据,利用遗传算法得到7个环境因素中影响比较大的环境因素是养殖点离国道的距离和养殖点离河流的距离等。可以归结为,国道上人流密集,容易造成鸭类传染病的区域性暴发;而河流的因素可能与候鸟的作用有关。所以,人类和候鸟活动都可能对鸭类传染病的发生起到很大的作用。

(3)现在的工作对建立鸭类传染病预测模型起到基础作用。然而,由于采样的科学性限制,在调查区的数据采样过程中,采样区不一致,有些地方样点多,有些地方样点少,结果没有那么理想。笔者的目标是将来通过数据调查和采样的深入,获取更好的数据,同时也希望把时间因素纳入进来,利用SIR(susceptible-infected-recovered)模型对鸭类进行传染病趋势的分析[10],建立更好的时空趋势模型,为鸭类传染病的防控做出贡献。

参考文献

[1] http://www.dnr.state.md.us/dnrnews/infocus/ai_faq.asp[accessed 29 March 2009].

[2] http://www.who.int/csr/disease/avian_influenza/avian_faqs/en/index.html[accessed 29 March 2009].

[3] GILBERT M, XIAO X M, CHATIAWEESUB P, et al. Avian influenza, domestic ducks and rice agriculture in Thailand [J]. Agric Ecosyst Environ, 2007,119:409 - 415.

[4] TRAPMAN P, MEESTER R, HEESTERBEEK H. A branching model for the spread of infectious animal diseases in varying environments [J]. Journal of Mathematical Biology, 2004,49:553.

[5] 崔尚金,王靖飞,吴春燕,等. 高致病性禽流感时空分布规律研究 - 传播的风险评估框架的初步建立[J]. 中国禽业导刊,2005,22(20):18.

[6] http://www.who.int/[accessed 1 November 2008].

[7] VOLLRATH I. Handling vague and qualitative criteria in case-based reasoning applications[C]. London, UK: Springer-Verlag, 2000:309 - 321.

[8] 齐平,贾瑞玉,贾兆红,等. 用遗传模拟退火算法挖掘特征项权重的研究[J]. 计算机技术与发展,2007,17(2):144.

[9] 贾兆红,陈华平. 基于改进遗传算法的权重发现技术[J]. 计算机工程,2007,33(5):156 - 157.

[10] GUAN Y, CHEN H, LI K S, et al. A model to control the epidemic of H5N1 influenza at the source BMC [J]. Infectious Diseases, 2007,7:132.