基于NIS的异常检测算法*)

徐 建 游 静 陈 昊 刘凤玉

(南京理工大学计算机科学与技术系 南京210094)

摘 要 该文根据生物免疫系统的免疫识别机理提出了一种基于NIS的异常检测算法来识别计算机系统运行的性能异常,将健康的系统状态作为"自我",不健康的系统状态作为"非我",多次应用阴性选择充当过滤器,并以遗传算法进化检测子,最后仿真实验验证了算法具备较好的检测性能。

关键词 生物免疫系统,阴性选择,遗传算法,异常检测

An Anomaly Detection Algorithm Based on NIS

XU Jian YOU Jing CHEN Hao LIU Feng-Yu

(Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094)

Abstract An anomaly detection algorithm which is based on NIS(Natural Immunity System)for the runtime performance exception of computer system is brought forward, in which healthy system states are looked as "self", unhealthy states are looked as "non-self". This algorithm applies NSA(Negative Selection Algorithm)time after time to act as filter and applies genetic algorithm to evolve the set of detector. Finally, the experimental results show that the proposed algorithm has a good detection effect.

Keywords Natural immunity system, Negative selection, Genetic algorithm, Anomaly detection

1 引言

近年来,人们对生物系统的许多有益特性在科学计算领域的应用产生了浓厚的兴趣,继人工神经网络和遗传算法被广泛研究和应用后,基于生物免疫系统的研究成为新的研究热点之一。生物免疫系统的核心功能是免疫识别,识别的本质就是区分"自我"和"非我"。免疫识别是通过生物体内淋巴细胞上的抗原识别受体与抗原的结合实现的,结合的强度称为亲和度。未成熟的 T 细胞首先要经历一个审查环节,只有那些不能与自我(即机体本身组织)发生免疫应答的 T 细胞才可以离开胸腺,执行免疫应答任务,从而防止免疫细胞对机体造成错误的攻击,这一过程为阴性选择(Negative Selection),这种免疫识别机理已经成功地应用在图像识别[9]、网络入侵检测[6,16]等领域。

本文根据生物体的免疫识别机理提出了一种基于NIS的异常检测算法来识别计算机系统运行的性能异常,将健康的系统状态作为"自我",不健康的系统状态作为"非我",也就是异常。从系统提取的性能属性如内存使用率、CPU的占用率、网络带宽、资源使用率、工作负荷、响应时间、服务时间等被作为基因对待,以某种方式编码生成染色体。可以有多种方式产生检测子,使用特定匹配规则的阴性选择作为审查过程,排除与"自我"发生绑定的检测子,直到产生数目足够的检测子为止,然后应用遗传算法对检测子集合进行进化操作,在进化过程中阴性选择充当了过滤器的角色,使得算法能够获得较高的检测性能。

2 相关的工作

很多异常检测算法都是通过从收集的数据集中学习到表示异常或正常的模型,在遇到异常的数据时产生警报,这种方式不被不平衡的数据集所影响,并且能够处理为曾见过的数据,已经在网络入侵检测[6,16]、疾病监控等领域取得很好的效果。

Philip Chan[13]提出了LERAD异常检测方式用于网络入侵检测,而后针对疾病监控的数据对数据处理方式作了修改,开发了疾病监控警报系统,已成功应用于波士顿的儿童医院。这种方式假设所有的训练数据都是正常的,通过学习得到逻辑规则用于异常检测,是个离线算法。为了消除所有训练数据都必须是正常的前提条件,他又提出了CLAD异常检测算法,该算法使用聚类方式识别局部的和全局的孤立点作为异常,它与其它聚类算法的一个显著不同之处在于每个簇有相同的宽度,簇与簇之间可以存在重叠部分。该算法是一个在线算法,不需要外在的训练阶段。Qiang Chen[14]提出了基于chi-square统计的多变元异常检测算法,该算法构造系统正常状态库,明显偏离正常状态的被识别为异常,在网络入侵检测方面取得较好的结果。Nguyen[15]提出了基于支持向量机的无监督异常检测算法来学习正常的系统状态向量,在检测阶段明显偏离支持区域的新的数据被标定为异常。

本文提出的基于生物免疫系统的异常检测算法是有监督的学习算法,它充分应用了生物体的免疫识别机理,通过把阴性选择多次充当过滤器,保证检测子的识别正确性,同时通过遗传算法对检测子进行进化操作,大大提高了检测子的多样性,能够更好地识别未曾见过的数据,仿真实验表明该算法有较好的性能。

3 算法实现

3.1 问题描述

本文要解决的一类问题其问题域由 n 个属性 $X_1, X_2, \cdots, X_n$ 所组成,这些属性的值既可以是离散的也可以是连续的。

*)本课题得到国家自然科学基金(No. 60273035)资助。徐 建 博士研究生,主要研究领域为软件自愈与抗衰,信息安全。游 静 博士研究生,主要研究领域为软件自愈与抗衰,信息安全。陈 昊 硕士研究生,主要研究领域为模式识别与人工智能,数据挖掘。刘凤玉 教授,博士生导师,主要研究领域为人工智能和网络安全。

实例是由"属性-值"对加上类标识表示的;学习到的 IF-THEN 规则的 THEN 部分具有离散的输出值,算法的核心任务就是要把样本准确地分类到相应的类别。

3.2 离散化连续值属性

从真实世界收集到数据集往往都是些连续的属性值,为了能够有效地处理这些数据,必须经过离散化这个预处理过程,以适应于算法的需要。现有的离散化算法有 Fayyad[4]提出的基于熵的多区间离散化算法;Kerber[10]提出用 ChiMerge 方式自底向上合并连续区间的离散化算法;Kohonen[11]提出了基于 LVQ(Learning Vector Quantization)的离散化方法等。本文采用了 Fayyad[4]提出的基于熵的多区间离散化算法实现连续属性值的离散化,它不需要事先确定区间的数目,能够满足本文检测算法的编码要求。

3.3 检测子的编码

在生物体中每一个染色体都是由若干个基因排列组合来决定的,而每一个基因又是由许多的核苷酸组成的,每一个核苷酸代表基因中的一位。在本文的算法中,每一个属性就表示一个基因,每一个属性离散化后的区间个数就代表核苷酸的数目,也表示基因的长度,检测子对应于染色体。举个例子来说,检测子的基因型和显型如图1所示,第一个属性由三个合法的值 TCP,UDP 和 ICMP. 分别代表着该基因的第一,二,三位,我们允许不止一位被置位,各位通过析取运算来组合。我们所得到每一个检测子都以分类规则的形式存在,而分类规则集合的自然表达形式是析取范式(DNF)的集合。每一条规则(检测子)的 IF-PART 是需要被检测的若干个属性的合取,THEN-PART 是赋予这条规则的类标号。

3.4 检测子的产生

有很多的方法在 non-self 空间中产生检测子,D'haeseleer[12]中提出了以随机的方式产生,通过阴性选择与 self 进行匹配来筛选,直到产生由概率分析计算得到的检测子数目。这种方法的缺点是计算复杂性会随着 self 集合大小成指数级增长,优点是能使用任意的匹配规则。特定的匹配规则要求特定的检测子产生方式与之相适应,由于本文匹配规则采用了任意的 M 个属性相匹配的方式,使得检测子编码串中的属性排列方式不影响检测结果,下面给出相应的构造检测子集合的方法,分两个阶段来进行,先产生原始的检测子集合,再经过一个一般化(generalization)形成最终的检测子,使得用比较少的检测子覆盖几乎全部异常空间。

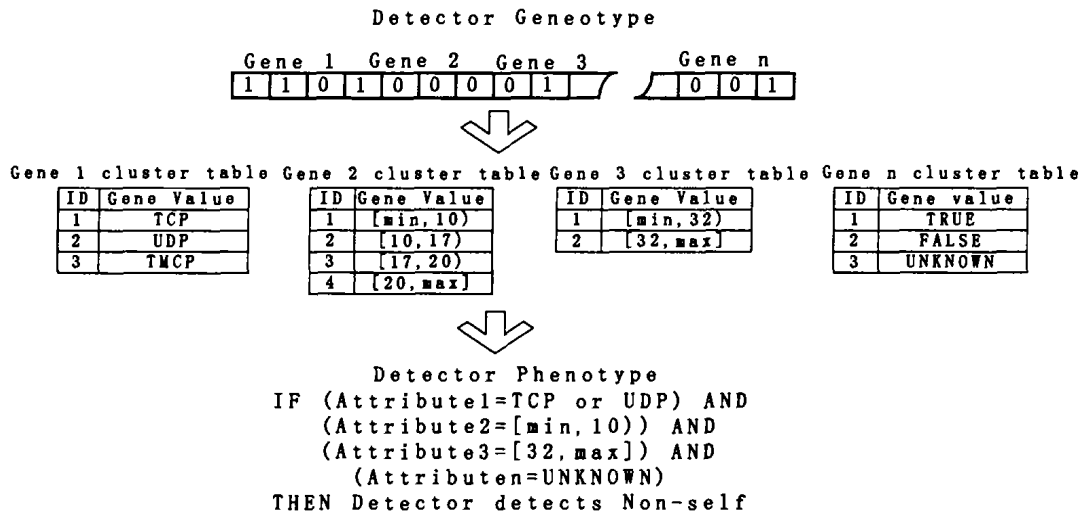


图1 Detector Geneotype and Phenotype

3.4.1 Primitive_D 的构造 本文构造检测子集合的方法,它基于这样的假设:正例 PE 和反例 NE 在属性集合下是一致的,满足 $PE \cup NE = E$(全集), $PE \cap NE = \Phi$。

```
Primitive_D Generate Algorithm
    Primitive_D=Empty.
    For each Ex training example
        IsPrimitive_D=True
        For each Ce counter-example to Ex
            Build D(Ex,Ce)
            If(D(Ex,Ce)==False)
                IsPrimitive_D=False
                Break
        If(IsPrimitive_D==True)
            Add Primitive-D(Ex).
```

Ce 对于正例来说是反例,对于反例来说正例;$D(Ex,Ce)$ 表示 Ex 和 Ce 匹配的基因(属性)数目最多不超过 M 则返回 TRUE,否则返回 FALSE;Add Primitive-D(Ex)表示把 Ex 加入到 Primitive_D 集合。Primitive_D 是利用阴性选择所生成的检测子集合。

3.4.2 final_D 的构造 final_D 是由 Primitive_D 的一般化(generalization)生成的。先给出 ε-Neighbor 含义,它代表两个任意检测子,如果不同的基因片断最多不超过 ε,则它们是 ε-Neighbor,ε 是预先定义的阈值。下面给出一般化描述:对于任意的 $d_1, d_2 \in$ Primitive_D,如果它们 ε-Neighbor,则一般化,从 Primitive_D 中移走 d_1 和 d_2,添加 $d_1 \vee d_2$ 进 final_D。最后把 Primitive_D 中剩余的检测子移到 final_D 中。通过一般化过程使得用少量的检测子可以覆盖几乎全部的异常空间。

3.5 进化

我们使用三个指标来衡量检测子集合的检测能力,即考虑检测子集合的完整性、一致性和简单性。

定义1 在测试样本集合 E 的正例集 PE 和反例集 NE 下,由 final_D 构成的逻辑规则 F,使得 $\forall e_j^- \in NE, F(e_j^-) \rightarrow$ True 或1,则称 F 所描述的检测子集合 final_D 在 E 上满足完整性(completeness);若对于 $\forall e_j^+ \in PE, F(e_j^+) \rightarrow$ False 或0,则称 F 所描述的检测子集合 D 在 E 上满足一致性(consistency)。那么,

(1)完整性的度量:

$$D_{TP}(F) = |F|/|NE| \tag{1}$$

其中,$|F|$ 为满足 $F(e_j^-) \rightarrow$True 或1的反例的数量;

(2)一致性的度量:

$$D_{FP}(F) = (|PE| - |F|)/|PE| \tag{2}$$

其中,$|F|$ 为满足 $F(e_j^+) \rightarrow$False 或0的正例的数量。

TP(True Positive)的含义是检测出的异常样本数目占实际的异常样本数目的比例,FP(False Positive)的含义是被误检成异常的正常样本的数目占正常样本数目的比例。

显然，D_{TP}，$D_{FP}\in[0,1]$、$D_{TP}=1$，$D_{FP}=0$分别表示在 E 上满足完整性和一致性。我们希望异常检测系统有较高的 TP 和较低的 FP，然而在真实世界中 TP 和 FP 的重要程度是有区别的，使 TP 尽可能地接近于100%，而容忍由此导致的 FP 的值相对较高是可以接受的，因此我们给 TP、FP 赋予不同的权值 w_{tp}，w_{fp}，来度量它们的重要程度。

定义2 对于给定的检测子集合 D，D 的简单性为 D 的基数与样本集合 E 的所有属性值域的基数和之比：

$$D_{SP}=|C|/\sum_{j=1}^{n}m_j \tag{3}$$

显然 D_{SP}越小表明 D 越简单，我们采用它的一种变型：

$$D_{SP}=(\sum_{j=1}^{n}m_j-|C|)/\sum_{j=1}^{n}m_j \tag{4}$$

综合以上几点，给出如下进化目标函数：

$$F(X,Y,Z)=w_{tp}X-w_{fp}Y+w_{sp}Z \tag{5}$$

其中变量 X,Y,Z 分别代表 D_{TP}，D_{FP}和 D_{SP}，w_{tp}，w_{fp}，$w_{sp}\in(0,1]$是它们的相对权重，$w_{tp}\geqslant w_{fp}$，进化目标为 F 取最大值 F_{max}。

进化过程采用遗传算法，其中选择算子采用稳态选择，最大程度地继承已获得的检测子，实现增量学习；交叉算子采用两点交叉，去除尾点效应；因为变异概率比较小，一些个体可能根本不发生一次变异，为了避免计算资源浪费，变异算子采用变通措施，首先进行个体层次的变异发生的概率判断，然后实施基因层次上的变异操作。

3.6 检测

检测的方式用过程 M-Belongs(E,final_D,M)来表示：

```
M-Belongs(E,final_D,M):
    NS=0
    For each detector d in the set of final_D
        For each attribute att_k in E
            IF att_k satisfies d_k
                NS=NS+1
                IF(NS≥M)
                    Return true
Return false.
```

其中 E 待检测的样本；final_D 检测子集合；d_k 是检测子的第 K 个基因，att_k satisfies d_k 意味着有相同的等位基因，M 是匹配阈值。

3.7 算法描述

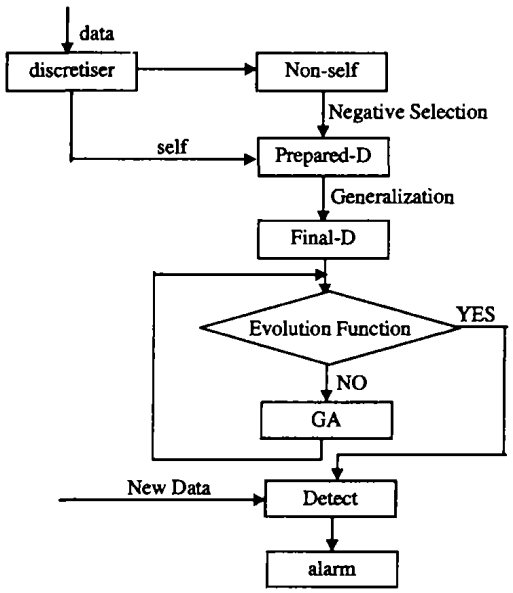


图2 AIS 算法流程图

本文的异常检测算法采用了 Smith[5]niching 策略，借鉴了 Kim[6]的检测子编码方式，采用了它的改进型。对于检测出异常情况，本文的算法与他们的算法相比主要有以下几个不同之处：1)采用了简单而有效的检测子二进制基因型、表现型。2)为了适应这种编码方式，采用基因的部分匹配方式来度量距离，为此引入了匹配阈值，建立了新的适应度评价函数。3)引入了泛化操作，减少了检测子的数目，显著地提高了 TP 的值，FP 的值控制在可接受的范围内。4)设计了进化目标函数，使系统朝着高 TP，低 FP 方向进化。5)阴性选择算子多次充当了过滤器的角色。图2给出了算法的流程图。

4 仿真实验

4.1 实验描述

本文采用了机器学习 UCI[17]数据库中的一组数据。数据是关于 Wisconsin 癌症，有32个属性，第一个为序列号，第二个为类标识，其它30个属性为监控属性，它们的值均为连续值。该数据集由569个样本组成，分成两类：212个恶性的(Malignant)和357个良性的(Benign)，我们把 Benign 定义为自我，而 Malignant 定义为非我，算法生成的检测子集合检测非我的，任何没有检测出的都认为是自我的。

采用了十重交叉验证[7]的方法从训练样本进化成检测子集合，用一个测试集合来检测未见过的样本。进化阶段每一代检测子选择 B 个最好的检测子通过遗传操作代替最差的 W 个，其中 B，W 的值每一代检测子的30%，交叉概率取0.6，变异概率取0.001。进化的终止条件是 TP 值为100%并且 FP 值为0%，如果不满足上述终止条件，则每一代最多进化100次终止。特征匹配阈值 M 是一个参数，在实验中根据真实情况下的权重调整它的值，确定最优值。实验中同时考虑样本数据的顺序，发现因为采用了十重交叉验证的方式，检测性能对样本出现顺序不敏感，检测结果基本没有变化。

4.2 实验结果

不同的匹配阈值 M 下的检测子集合的不同的检测结果如表1所示。

表1 不同匹配阈值 M 下的 TP，FP，D 的值

M	TP	FP	AD
12	91.42±1.03%	3.50±0.42%	10.2
13	93.16±0.24%	3.62±0.60%	24.2
14	97.40±0.71%	7.00±0.89%	51.5
15	98.58±0.10%	10.42±0.22%	61.0
16	99.06±0.12%	10.64±0.31%	72.4
17	98.82±0.24%	9.94±0.78%	64.1
18	99.32±0.26%	10.13±0.53%	52.5
19	99.54±0.46%	8.22±0.46%	50.2
20	99.33±0.23%	5.74±0.45%	46.2
21	99.31±0.22%	5.88±0.30%	39.5
22	98.58±0.10%	5.04±0.28%	31.8
23	97.80±0.63%	3.36±0.10%	29.7
24	97.04±0.51%	4.76±0.42%	24.3
25	96.70±0.47%	4.76±0.30%	19.1
26	97.48±0.16%	2.33±0.47%	16.0
27	95.04±0.70%	2.94±0.14%	11.5

从表1中的数据我们可以看出，TP 大部分的值都高于98.0%，FP 大部分的值都低于5.0%，算法检测效果是比较好的，在 $M=19$时，能完全检测出异常情况，虽然 FP 的值也相对较高，在 w_{tp}比 w_{fp}大很多时也是可以接受的。我们可以针对特定的问题而给出的权重 w_{tp}，w_{fp}，w_{sp}来选择匹配阈值从而使目标函数值最大。UCI 提供的在这组数据集上所获得的最好的分类精度为97.5%。

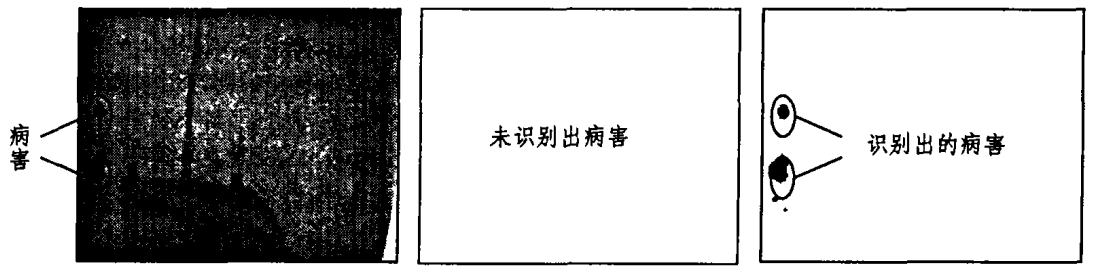
结论和展望 本文提出的算法吸收了来自生物免疫系统免疫识别的灵感，采用了改进的基因编码方式，并且采用了新的部分匹配方式来适应这种编码方式，通过泛化操作减少检测子的数目，同时覆盖尽可能大的异常空间，最后经过遗传算法进化检测子，并保证检测子的多样性，仿真实验结果表明该算法具备较好的检测性能。同时实验结果也表明检测性能对参数比较敏感，需要一个参数训练的过程；检测性能对数据样本的表示顺序不敏感。在真实系统的状态属性集合中，并不是每一个属性对于系统性能的作用都是相同的，本算法没有根据属性的重要程度赋于相应的权重，而假设每一个属性有相同的地位，未来的异常检测算法可以考虑这一方面对于检测性能的影响。

参 考 文 献

1 Forrest S, Hofmeyr S A. Immunology as information processing [A]. In: Segel L. A, Cohen I. R, eds. Design Principles for the Immune System and Other Distributed Autonomous Systems [C]. USA: Oxford University Press, 2000
2 Forrest S, et al. Self-Nonself Discrimination in a Computer. In: Proc. of 1994 IEEE Symposium on Research in Security and Privacy, Los Alamos, CA: IEEE Computer Society Press, 1994
3 D'haeseleer P. A Distributed Approach to Anomaly Detection. ACM Transactions on Information System Security, 1997
4 Fayyad U M, Irani K B. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In: Proc. of The Thirteenth intl. Joint Conf. on Artifical Intelligence, 1993. 1022～1027
5 Smith R E, et al. Searching for Diverse, Coperative Populations With Genetic Algorithm [J]. Evolutionary Computation, 1997, 1 (2): 127～149
6 Kim J, Bentley P J. Towards an Artificial Immune System for Network Intrusion Detection: An Investigation of Clonal Selection with a Negative Selection Operator. the Congress on Evolutionary Computation (CEC-2001), Seoul, Korea, May 2001. 1244～1252
7 Written I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers
8 Kephart J O, Chess D M. The vision of autonomatic computing. Computer, 2003, 36(1): 41～52
9 McCoy D F, Devarajan V. Artificial immune systems and aerial image segmentation. In: Proc. IEEE Intl. Conf. on Systems, Man, and Cybemetics, Orlando, Florida, 1997. 867～872
10 Kerber R. ChiMerge: Discretization of Numeric Attributes, Learning: Inductive, AAAI92, 1992. 123～128
11 Kohonen T. Self-Organizing Maps. Springer Verlag, 1995
12 D'haeseleer P, Forrest S, Helman P. An Immunological Approach to Change Detection: Algorithms, Analysis and Implications. IEEE Symposium on Security and Privacy, 1996
13 Matthew V, Mahoney, Philip K. Chan: Learning Rules for Anomaly Detection of Hostile Network Traffic. In: Proc. Third IEEE Intl. Conf. on Data Mining (ICDM), 2003. 601～604
14 Ye Nong, Chen Qiang. An Anomaly Detection Technique Based on A Chi-square Statistic for Detecting Intrusions into Information Systems, Quality and Reliability Engineering International, 2001, 17(2): 105～112
15 Nguyen B V. An Application of Support Vector Machines to Anomaly Detection, CS681 (Research in Computer Science - Support Vector Machine)report, Fall 2002
16 李千目，张琨，等. 一种基于生物免疫学的入侵检测系统. 计算机工程与应用，2003, 39(8): 45～48
17 Murphy P M, Aha D W. UCI Repository of machine learning databases, 1992

（上接第101页）

中使用本文提出的二值化方法比使用直方图最频法在误检率不变的基础上有效降低了破损识别的漏检率，提高了破损识别的正确率。实验结果如表1所示。



(a) 原始图像　(b) 使用直方图最频法去除车影干扰的识别效果图　(c) 使用基于种子填充的二值化方法去除车影干扰的识别效果图

图3 两种方法的最终识别效果

表1 使用两种方法去除检测车影后的破损识别效果

	误检率	漏检率	正确率
直方图最频法	1%	11.5%	87.5%
基于种子填充的图像二值化方法	1%	3%	96%

从表1中可以看出，本文提出的基于种子填充的图像二值化方法在漏检率和正确率方面都比直方图最频法要好。

结束语 在公路路面病害检测的去除检测车影的应用中，由于目标和背景的灰度对比比较明显，因此采用了将直方图最频法和漫水法相结合的基于种子填充的图像二值化方法。由前面的介绍可以看出阈值化方法其实多种多样，因此根据不同的应用需要，可以将种子填充方法应用到不同的阈值化方法中来提高阈值化的精度。此外种子填充中初始种子点的选择方法还需要改进，在本文给出的交互式指定方法自动程度不高，使用时不够方便，而模板匹配方法比较复杂，不够直接，是否能找到一种不需要人工干预，又简洁明了的确定初始种子点的方法，这还需要进一步的研究。

参 考 文 献

1 Pynn J, Wright A, Lodge R. Automatic Identification of Cracks in Road surfaces. In: Proc. of 7thIntl. Congress on Image Processing and its Applications, IEE, 1999, 2: 671～675
2 Wang K C P, Elliott R P. Investigation of Image Archiving for Pavement Surface Distress Survey: A final report submitted to Mack-Blackwell Transportation Center. July, 1999
3 Paterson, William D. Proposal of Universal Cracking Indicator for Pavements: Transportation Research Record, Washington, D. C., 1994, 1455: 69～76
4 李晋惠，楼伟，姜寿山. 基于CCD的公路路面病害检测技术研究. 西安工业学院学报，2002, 22(2): 95～99
5 李冠. 灰度文档图像的直接局域二值化方法：[南开大学硕士研究生毕业论文]. 天津，2002
6 Yang H S. Split-and-Merge Segmentation Employing Thresholding Technique. Image Processing. In: Proc. Intl. Conf. on Published, 1997, 1: 239～242
7 Rosin P L, Ioannidis E. Evaluation of global image thresholding for change detection. Pattern Recognition Letters, 2003, 24(14): 2345～2356
8 Mehnert A, Jackway P. An improved seeded region growing algorithm. Pattern Recognition Letters, 1997, 18(3): 1065～1071
9 刘相滨，胡峰松，张邦基. 一种新的区域种子填充算法. 计算机工程与应用，2002, 8
10 Prokop R J, Reeves A P. A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP: Graphical Models and Image Processing, 1992, 54(5): 438～460