Mathematical modeling

第十一讲 马尔科夫模型

周毓明

zhouyuming@nju.edu.cn

南京大学计算机科学与技术系

课程内容

1. 数学概念与模型
2. 实际案例与分析
3. 计算机典型应用

1. 数学概念与模型

马氏链模型

描述一类重要的随机动态系统（过程）的模型

- 系统在每个时期所处的状态是随机的
- 从一时期到下时期的状态按一定概率转移
- 下时期状态只取决于本时期状态和转移概率

  已知现在，将来与过去无关（无后效性）

马氏链 (Markov Chain)

——时间、状态均为离散的随机转移过程

健康与疾病

通过有实际背景的例子介绍马氏链的基本概念和性质

人的健康状态随着时间的推移会随机地发生转变

保险公司要对投保人未来的健康状态作出估计, 以制订保险金和理赔金的数额

例1. 人的健康状况分为健康和疾病两种状态，设对特定年龄段的人，今年健康、明年保持健康状态的概率为0.8, 而今年患病、明年转为健康状态的概率为0.7，

若某人投保时健康, 问10年后他仍处于健康状态的概率

健康与疾病

状态与状态转移

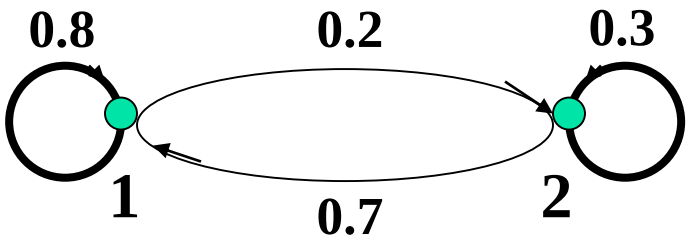
$$\text{状态}X_n=\begin{cases}1, & \text{第}n\text{年健康}\\ 2, & \text{第}n\text{年疾病}\end{cases}$$

状态概率$a_i(n)=P(X_n=i),\quad i=1,2,\ n=0,1,\cdots$

转移概率$p_{ij}=P(X_{n+1}=j\,|\,X_n=i),\ \ i,j=1,2,\ \ n=0,1,\cdots$

$$p_{11}=0.8\quad p_{12}=1-p_{11}=0.2$$

$$p_{21}=0.7\quad p_{22}=1-p_{21}=0.3$$



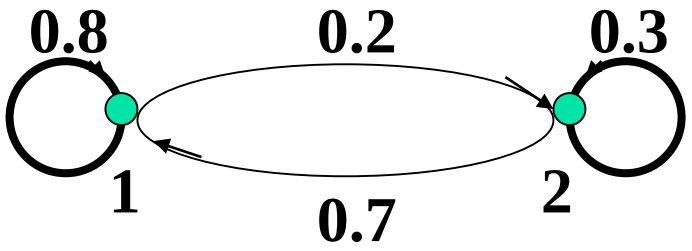
X_{n+1}只取决于X_n和p_{ij}, 与X_{n-1}, …无关

状态转移具有无后效性

$$a_1(n+1)=a_1(n)p_{11}+a_2(n)p_{21}$$

$$a_2(n+1)=a_1(n)p_{12}+a_2(n)p_{22}$$

健康与疾病



状态与状态转移

$$\begin{cases} a_1(n+1) = a_1(n)p_{11} + a_2(n)p_{21} \\ a_2(n+1) = a_1(n)p_{12} + a_2(n)p_{22} \end{cases}$$

给定$a(0)$, 预测 $a(n)$, n=1,2…

设投保时健康

n	0	1	2	3	…	∞
$a_1(n)$	1	0.8	0.78	0.778	…	7/9
$a_2(n)$	0	0.2	0.22	0.222	…	2/9

设投保时疾病

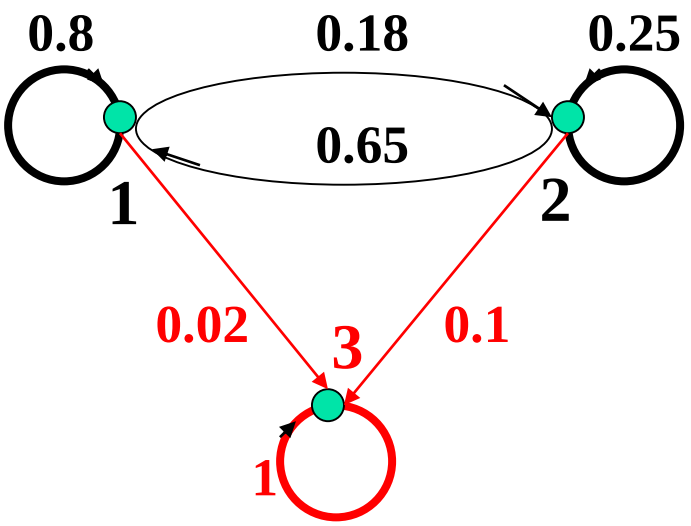
$a_1(n)$	0	0.7	0.77	0.777	…	7/9
$a_2(n)$	1	0.3	0.33	0.333	…	2/9

$n\to\infty$时状态概率趋于稳定值，稳定值与初始状态无关

健康与疾病

例2. 健康和疾病状态同上，X_n=1~ 健康，X_n=2~ 疾病

死亡为第3种状态，记X_n=3



p_{11}=0.8, p_{12}=0.18, p_{13}=0.02

p_{21}=0.65, p_{22}=0.25, p_{23}=0.1

p_{31}=0, p_{32}=0, p_{33}=1

$$a_1(n+1) = a_1(n)p_{11} + a_2(n)p_{21} + a_3(n)p_{31}$$
$$a_2(n+1) = a_1(n)p_{12} + a_2(n)p_{22} + a_3(n)p_{32}$$
$$a_3(n+1) = a_1(n)p_{13} + a_2(n)p_{23} + a_3(n)p_{33}$$

健康与疾病

状态与状态转移

设投保时处于健康状态，预测 $a(n)$, n=1,2…

n	0	1	2	3	…	50	…	∞
$a_1(n)$	1	0.8	0.757	0.7285	…	0.1293	…	0
$a_2(n)$	0	0.18	0.189	0.1835	…	0.0326	…	0
$a_3(n)$	0	0.02	0.054	0.0880	…	0.8381	…	1

- 不论初始状态如何，最终都要转到状态3；
- 一旦$a_1(k)= a_2(k)=0, a_3(k)=1$, 则对于$n>k$, $a_1(n)=0$， $a_2(n)=0$, $a_3(n)=1$, 即从状态3不会转移到其它状态。

马氏链的基本方程

状态 $X_n = 1,2,\cdots k \quad (n = 0,1,\cdots)$

状态概率 $a_i(n) = P(X_n = i)$,

$i = 1,2,\cdots,k,\ n = 0,1,\cdots$

$$\sum_{i=1}^{k} a_i(n) = 1$$

转移概率 $p_{ij} = P(X_{n+1} = j \mid X_n = i)$

$$p_{ij} \geq 0, \sum_{j=1}^{k} p_{ij} = 1, i = 1,2,\cdots,k$$

基本方程

$$a_i(n+1) = \sum_{j=1}^{k} a_j(n) p_{ji}, \quad i = 1,2,\cdots,k$$

$a(n) = (a_1(n), a_2(n), \cdots, a_k(n))$

~ 状态概率向量

$P = \{p_{ij}\}_{k\times k}$ ~ 转移概率矩阵

（非负，行和为1）

$$\mathbf{a(n+1) = a(n)P}$$

$$\mathbf{a(n) = a(0)P^n}$$

马氏链的两个重要类型

$$\mathbf{a(n+1)=a(n)P}$$

1. 正则链 ~ 从任一状态出发经有限次转移能以正概率到达另外任一状态（如例1）。

$$\text{正则链} \Leftrightarrow \exists N, P^N > 0$$

$$\text{正则链} \Rightarrow \exists w, a(n) \to w(n \to \infty)$$

w ~ 稳态概率

w满足 $wP = w$

例1. $P = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}$

$$\begin{cases} 0.8w_1 + 0.7w_2 = w_1 \\ 0.2w_1 + 0.3w_2 = w_2 \end{cases} \Rightarrow 0.2w_1 = 0.7w_2$$

w满足 $\sum_{i=1}^{k} w_i = 1$

$$w_1 + w_2 = 1 \Rightarrow w = (7/9, 2/9)$$

健康与疾病

马氏链的两个重要类型

2. 吸收链 ~ 存在吸收状态（一旦到达就不会离开的状态 i, $p_{ii}=1$），且从任一非吸收状态出发经有限次转移能以正概率到达吸收状态（如例2）。

有r个吸收状态的吸收链的转移概率阵标准形式

$$P=\begin{bmatrix} I_{r\times r} & 0 \\ R & Q \end{bmatrix}$$

R有非零元素

$$M=(I-Q)^{-1}=\sum_{s=0}^{\infty}Q^{s}$$

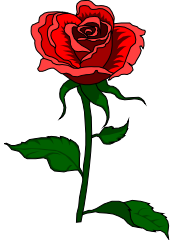
$$y=(y_1,y_2,\cdots y_{k-r})=Me$$

$$e=(1,1,\cdots,1)^{T}$$

y_i ~ 从第 i 个非吸收状态出发，被某个吸收状态吸收前的平均转移次数。

2. 实际案例与分析

基因遗传



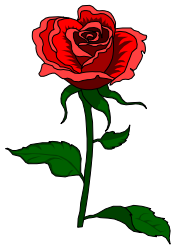
背景

- 生物的外部表征由内部相应的基因决定。
- 基因分优势基因***d*** 和劣势基因***r*** 两种。

完全优势基因遗传

- 每种外部表征由两个基因决定，每个基因可以是***d, r*** 中的任一个。形成3种基因类型：***dd*** ~ 优种***D***, ***dr*** ~ 混种***H***, ***rr*** ~ 劣种***R***。
- 基因类型为优种和混种, 外部表征呈优势；基因类型为劣种，外部表征呈劣势。
- 生物繁殖时后代随机地（等概率地）继承父、母的各一个基因，形成它的两个基因。父母的基因类型决定后代基因类型的概率

完全优势基因遗传



3种基因类型：*dd*~优种*D*, *dr*~混种*H*, *rr*~劣种*R*

父母基因类型决定后代各种基因类型的概率

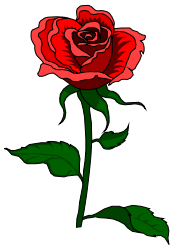
父母基因类型组合		*DD*	*RR*	*DH*	*DR*	*HH*	*HR*
后代各种基因类型的概率	*D*	1	0	1 / 2	0	1 / 4	0
	H	0	0	1 / 2	1	1 / 2	1 / 2
	R	0	1	0	0	1 / 4	1 / 2

$$P(D \mid DH)=P(dd \mid dd,dr)=P(d \mid dd)P(d \mid dr) \quad =1\times 1/2=1/2$$

$$P(R \mid HH)=P(rr \mid dr,dr)=P(r \mid dr)P(r \mid dr) \quad =1/2\times 1/2=1/4$$

随机繁殖

讨论基因类型的演变情况



假设

- 设群体中雄性、雌性的比例相等，基因类型的分布相同（记作$D:H:R$）

- 每一雄性个体以$D:H:R$的概率与一雌性个体配对繁殖，其后代随机地继承它们的各一个基因

- 设初始一代基因类型比例$D:H:R = a:2b:c$（$a+2b+c=1$), 记$p=a+b$, $q=b+c$, 则群体中优势基因和劣势基因比例 $d:r=p:q$ ($p+q=1$)。

建模

状态X_n=1,2,3 ~ 第n代的一个体属于D, H, R

状态概率 $a_i(n)$ ~ 第n代的一个体属于状态i(=1,2,3)的概率。

随机繁殖

状态转移概率

基因比例 ***d:r=p:q***

$$p_{ij}=P(X_{n+1}=j(\text{后代基因类型})\,|\,X_n=i(\text{父基因类型}))$$

$$p_{11}=P(X_{n+1}=1(\text{后代为}dd)\,|\,X_n=1(\text{父为}dd))=p$$

$$p_{12}=P(X_{n+1}=2(\text{后代为}dr)\,|\,X_n=1(\text{父为}dd))=q$$

$$p_{13}=P(X_{n+1}=3(\text{后代为}rr)\,|\,X_n=1(\text{父为}dd))=0$$

$$p_{21}=P(X_{n+1}=1(\text{后代为}dd)\,|\,X_n=2(\text{父为}dr))=1/2\cdot p=p/2$$

$$p_{22}=P(X_{n+1}=2(\text{后代为}dr)\,|\,X_n=2(\text{父为}dr))$$
$$=1/2\cdot p+1/2\cdot q=1/2$$

转移概率矩阵

$$P=\begin{bmatrix} p & q & 0 \\ p/2 & 1/2 & q/2 \\ 0 & p & q \end{bmatrix}$$

随机繁殖

马氏链模型 $a(n+1)=a(n)P, n=0,1,\cdots$

$$a(0)=(a,2b,c)$$

$$a(1)=a(0)P=(p^2,2pq,q^2)$$

$$a(2)=a(1)P=(p^2,2pq,q^2)$$

……

$$P=\begin{bmatrix} p & q & 0 \\ p/2 & 1/2 & q/2 \\ 0 & p & q \end{bmatrix}$$

$$p=a+b,\ q=b+c$$

$$a+2b+c=1$$

$a(0)$任意，稳态分布 $w=wP=(p^2,2pq,q^2)$

自然界中通常$p=q=1/2$　　稳态分布$D:H:R=1/4:1/2:1/4$

解释“豆科植物的茎，绿色:黄色=3:1”

基因类型为D和H, 优势表征——绿色，

基因类型为R, 劣势表征——黄色。

$(D+H):R=3:1$

3. 计算机典型应用

① 缺陷数目预测

② Pagerank算法

③ 其他应用…

Predicting Defect Numbers Based on Defect State Transition Models

Jue Wang and Hongyu Zhang
School of Software, Tsinghua University
Tsinghua National Laboratory for Information Science and Technology (TNList)
Beijing 100084, China
cecilia.juewang@gmail.com, hongyu@tsinghua.edu.cn

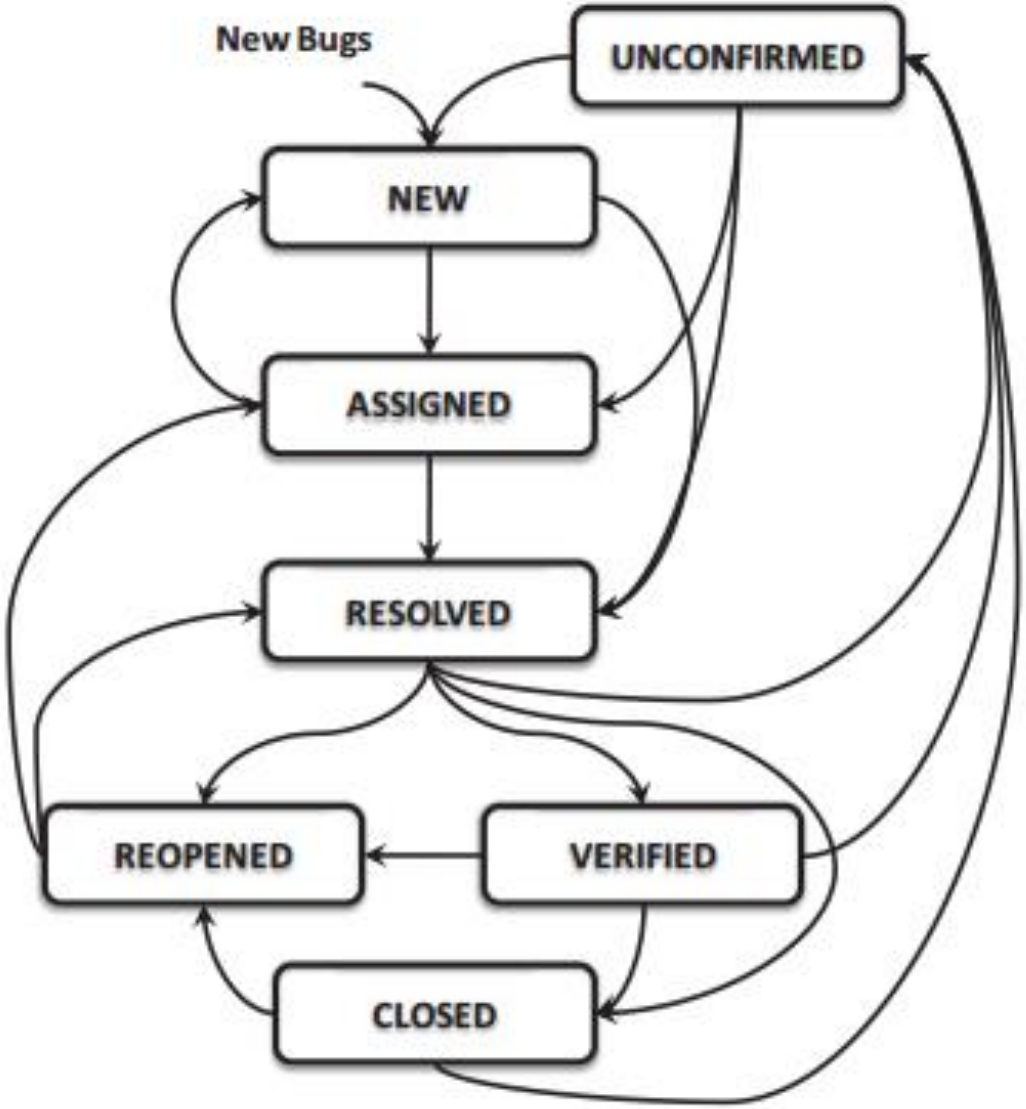


Figure 1. The defect state transition process of Bugzilla

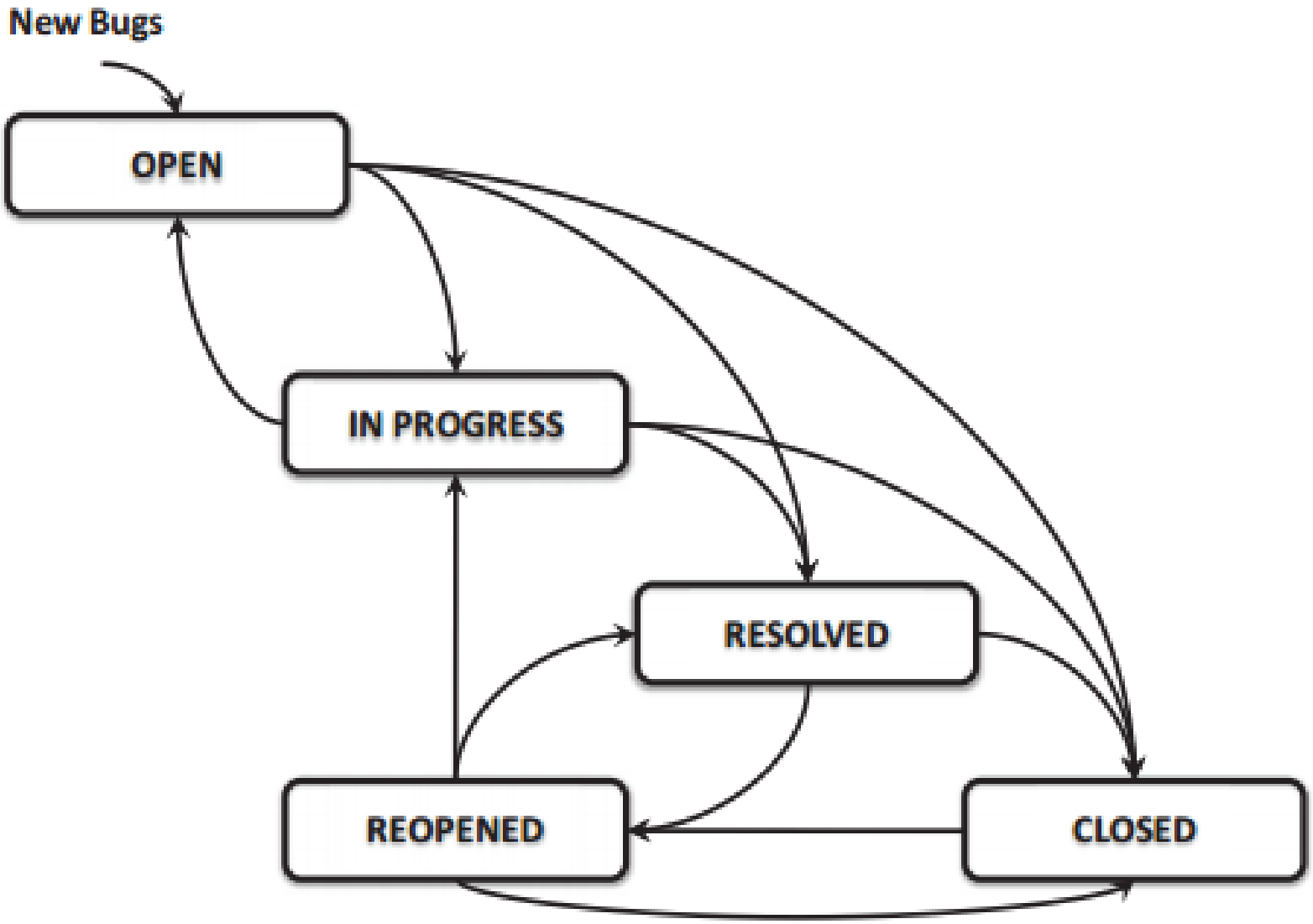


Figure 2. The default defect state transition process of Jira

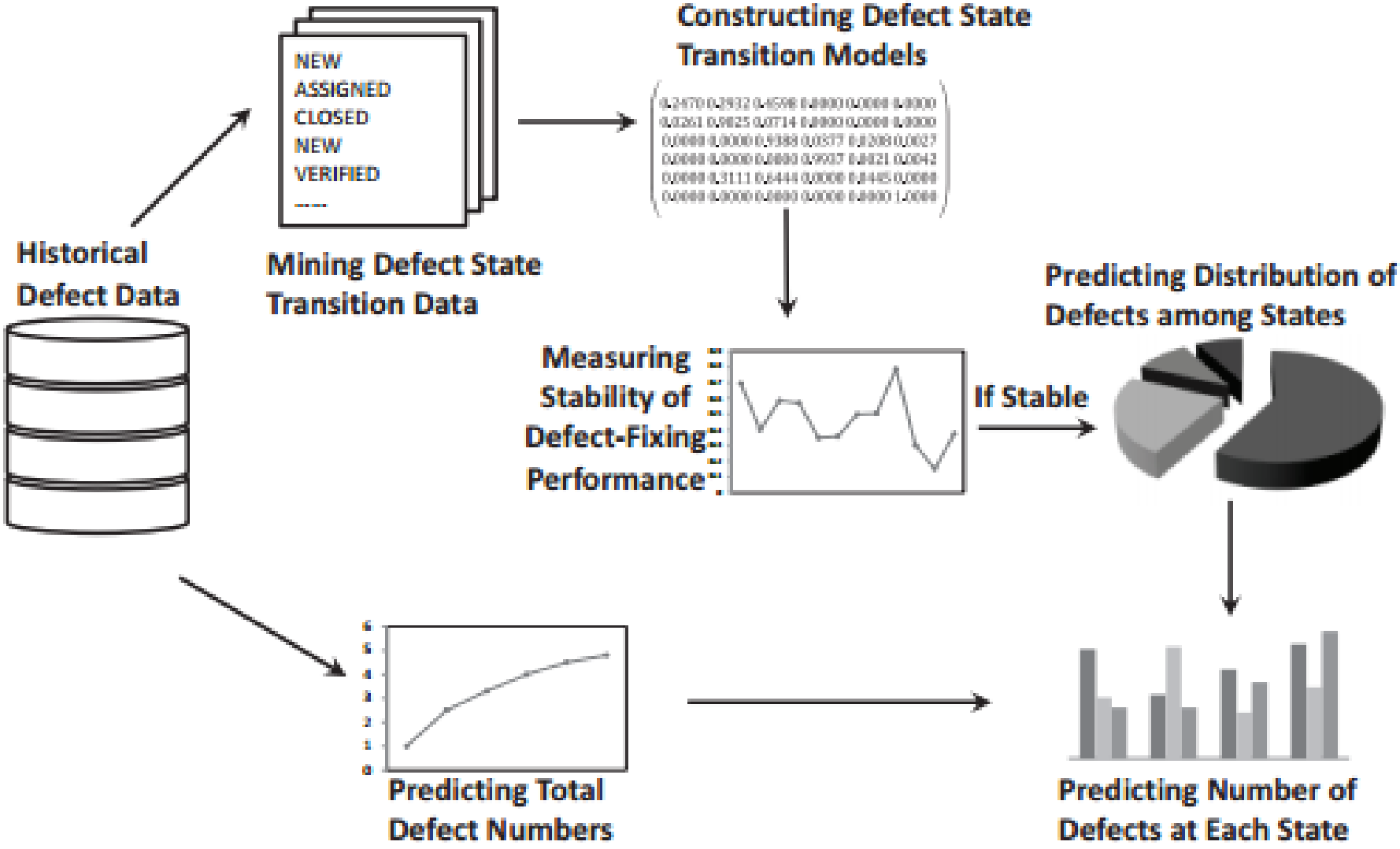


Figure 3. The overall structure of BugStates

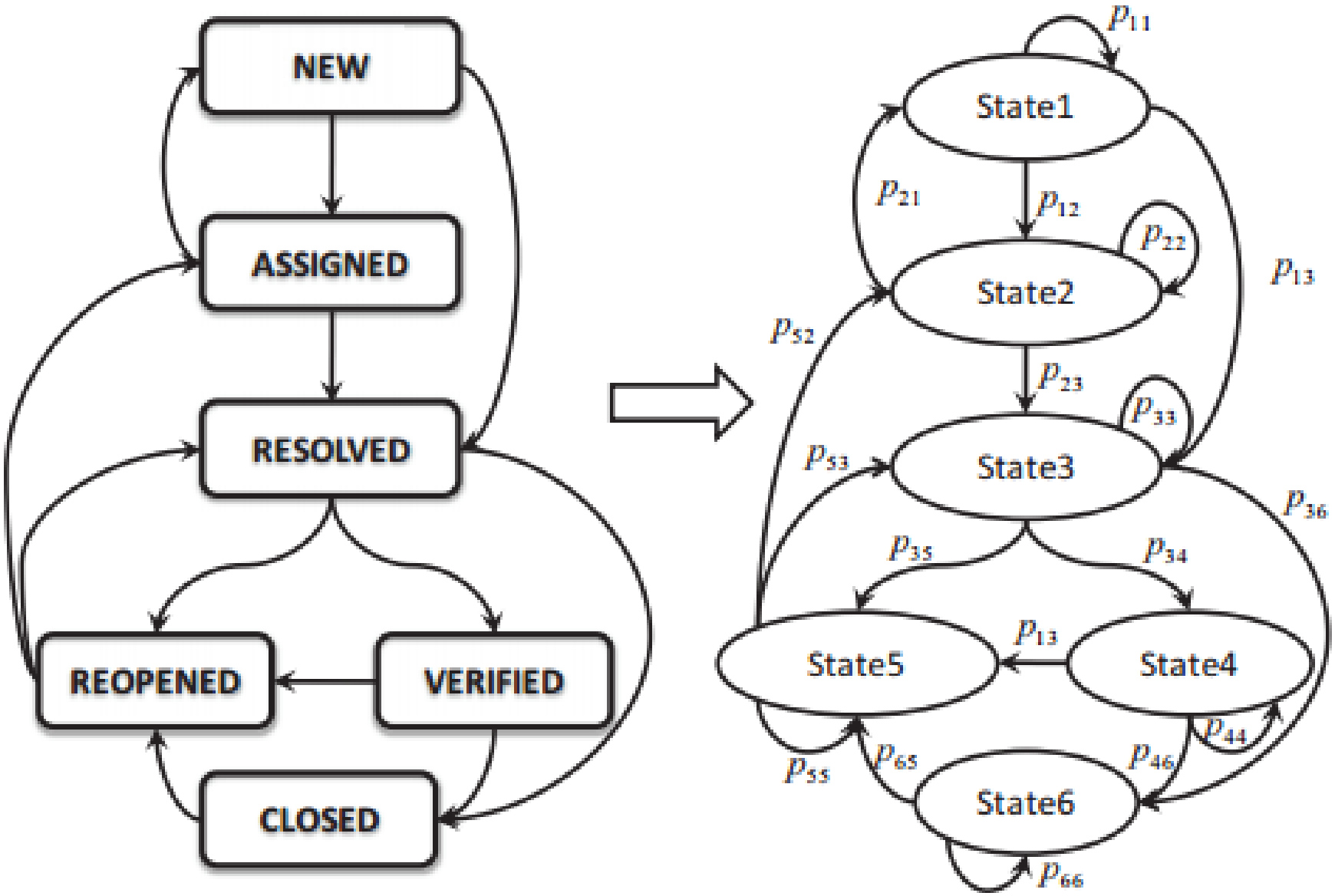


Figure 4. Mapping from Bugzilla defect state transition graph to Markov-like model

Table 1. The activity log of defect #280333[1] for Eclipse JDT.UI

Who	When	What	Removed	Added
pwebs-ter	2009-06-16 07:40:50	CC		pwebs-ter
markus_keller	2009-06-16 09:27:01	CC		markus_keller
		Status	**NEW**	**RESOLVED**
		Resolution		WONTFIX
paules	2009-06-22 12:09:30	**Status**	**RESOLVED**	**REOPENED**
		Resolution	WONTFIX	
daniel_megert	2009-06-23 02:30:39	**Status**	**REOPENED**	**NEW**
		…	…	…
markus_keller	2009-08-03 07:59:13	**Status**	**NEW**	**RESOLVED**
		Resolution		FIXED
daniel_megert	2009-08-06 03:26:16	**Status**	**RESOLVED**	**VERIFIED**
paules	2009-08-10 08:24:07	**Status**	**VERIFIED**	**CLOSED**

Table 2. The activity log of defect #2060[2] for Lucene-Java

Michael McCandless made changes – 14/Nov/09 11:22		
Field	Original Value	New Value
Status	**Open**	**Resolved** (Resolution Fixed)
Uwe Schindler made changes – 25/Nov/09 16:47		
Status	**Resolved**	**Closed**
Michael McCandless made changes – 30/May/10 12:40		
Status	**Closed**	**Reopened**
	...	
Uwe Schindler made changes – 18/Jun/10 08:03		
Status	**Resolved**	**Closed**

Table 3. The projects studied

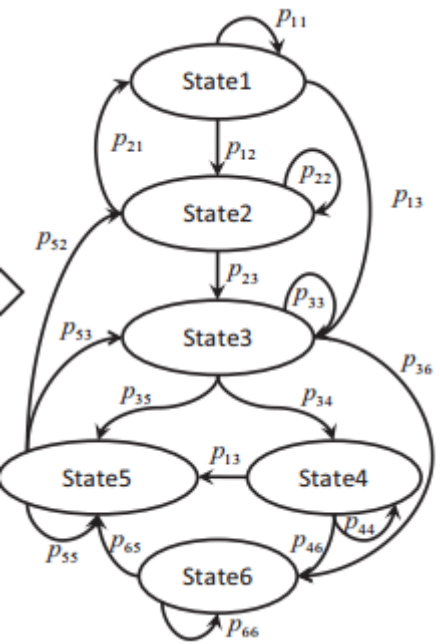
Project	Description	Defects Reported in 2009	Defects Reported in 2010
PDE.Build[3]	build support for Java	227	139
JDT.Text[4]	Java editing support	248	235
JDT.UI[5]	user interface for the Eclipse Java IDE	663	472
Platform. Debug[6]	debug support for Eclipse platform	348	231
Lucene-Java.Index[7]	the index component of the Apache Java search engine	47	65
Spring.NET[8]	a port and extension of the Java based Spring Framework for .NET	66	39

根据2009年的12个月数据构造变迁矩阵

$$P_{JDT.UI} = \begin{array}{c} \begin{matrix} NEW & ASSI. & RESO. & VERI. & REOP. & CLOSED \end{matrix} \\ \begin{pmatrix} 0.2470 & 0.2932 & 0.4598 & 0.0000 & 0.0000 & 0.0000 \\ 0.0261 & 0.9025 & 0.0714 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.9388 & 0.0377 & 0.0208 & 0.0027 \\ 0.0000 & 0.0000 & 0.0000 & 0.9937 & 0.0021 & 0.0042 \\ 0.0000 & 0.3111 & 0.6444 & 0.0000 & 0.0445 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix} \end{array}$$

在2009年12月31日时的状态分布

	NEW	ASSI.	RESO.	VERI.	REOP.	CLOS.
PDE.Build	(0.3744	0.0044	0.5991	0.0088	0.0000	0.0132)
JDT.UI	(0.0799	0.2157	0.5189	0.1237	0.0015	0.0603)
JDT.Text	(0.0363	0.2472	0.4960	0.1250	0.0040	0.0645)
Platform. Debug	(0.3420	0.0344	0.2414	0.3420	0.0115	0.0287)
Spring.NET	(0.4394	0.0303	0.5152	N/A	0.0000	0.0152)



根据2009年的12个月数据构造变迁矩阵

在2009年12月31日时的状态分布

得到2010年6月的状态分布

	NEW	ASSI.	RESO.	VERI.	REOP.	CLOS.
PDE.Build	(0.2792	0.0071	0.6643	0.0106	0.0000	0.0389)
JDT.UI	(0.0152	0.1967	0.5034	0.1751	0.0108	0.0988)
JDT.Text	(0.0093	0.2437	0.4506	0.1742	0.0132	0.1090)
Platform. Debug	(0.2340	0.0289	0.2525	0.4040	0.0176	0.0630)
Spring.NET	(0.3238	0.0697	0.5220	N/A	0.0121	0.0723)

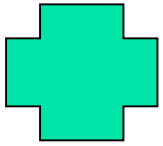
根据可靠性增长模型得到2010年6月各项目上的总bug数目

$$m_t = a(1-e^{-bt}),$$

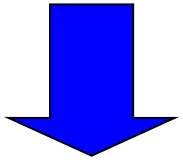
Table 5. The G-O models constructed in Experiment A

	a	*b*	R^2	#Actu.	#Pred.	MRE
PDE. Build	354.2	0.0890	0.990	311	283	9.00%
JDT.UI	1572	0.0450	0.997	905	875	3.31%
JDT.Text	383	0.0860	0.997	350	302	13.71%
Platform. Debug	978	0.0350	0.988	509	460	9.63%
Spring. NET	9455	0.0004	0.968	88	83	5.68%

根据可靠性增长模型得到2010年6月
各项目上的总bug数目

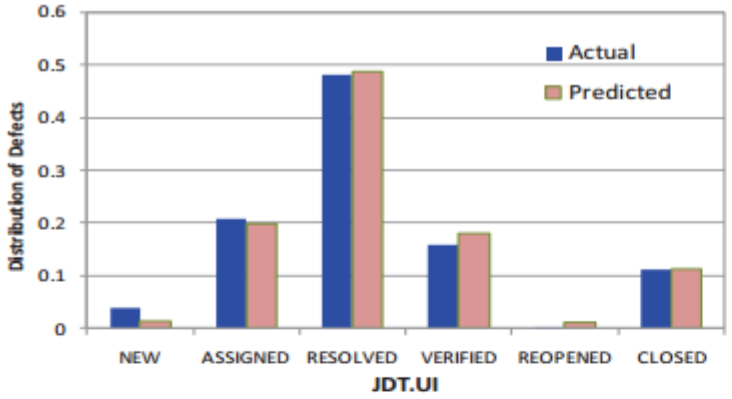


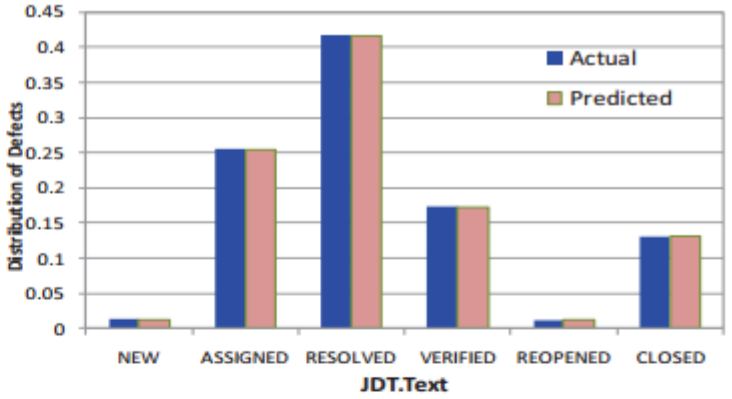
得到2010年6月的状态分布

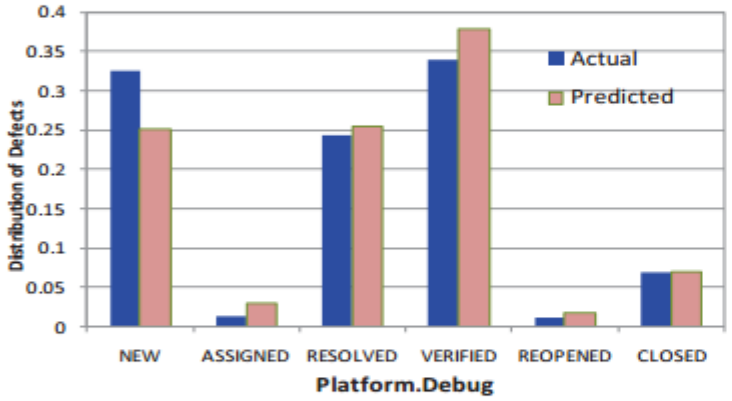


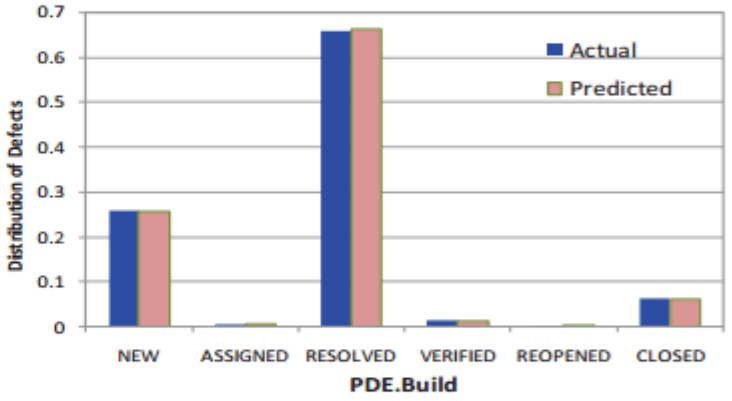
得到2010年6月的各状态上的bug数目

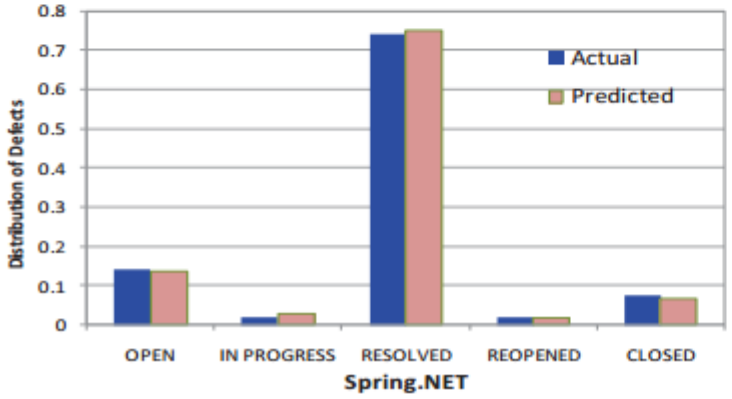
实际的状态分布 vs. 预测的状态分布

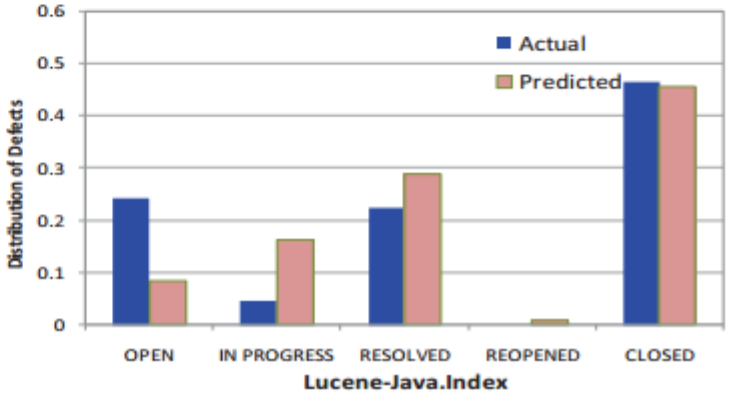




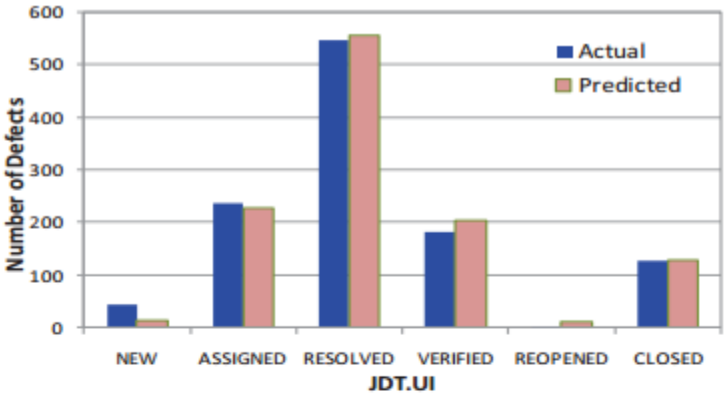


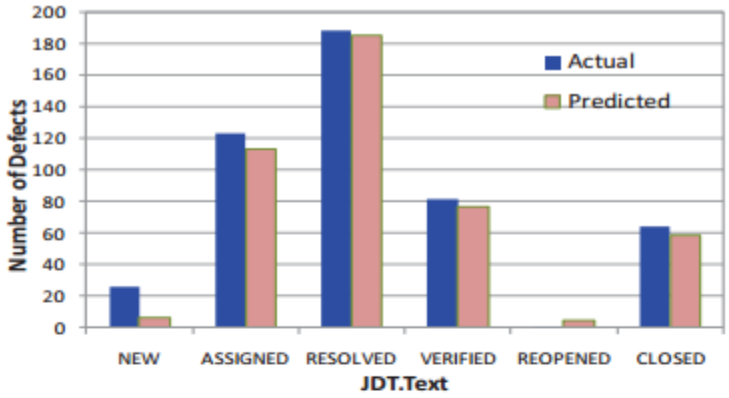


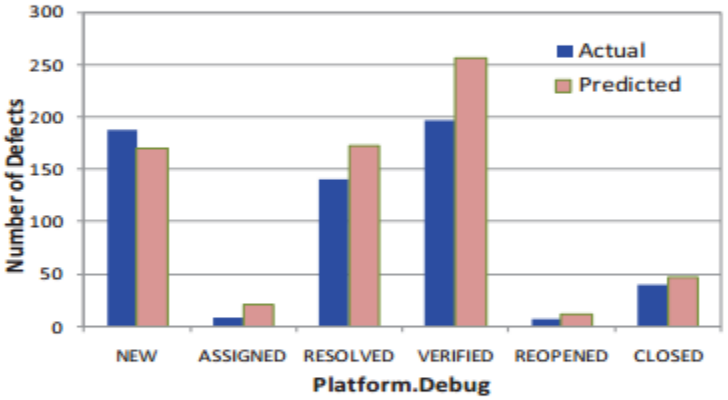


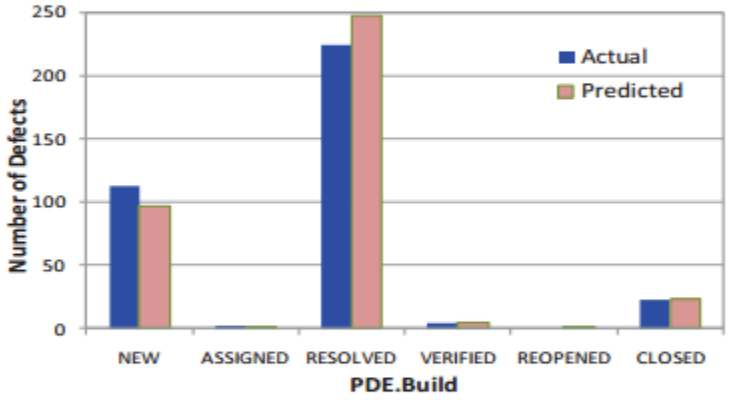


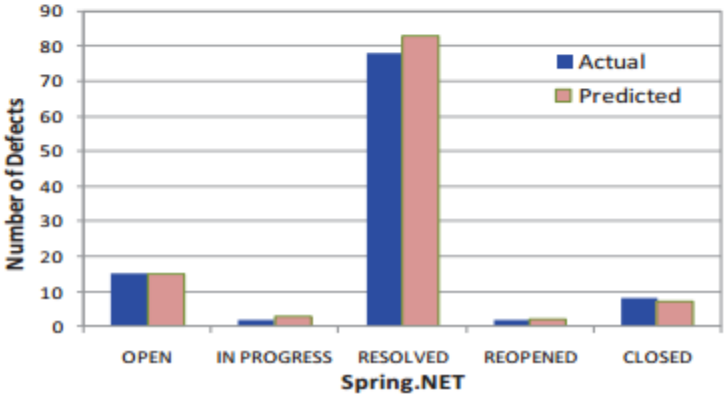
实际的bug数目 vs. 预测的bug数目

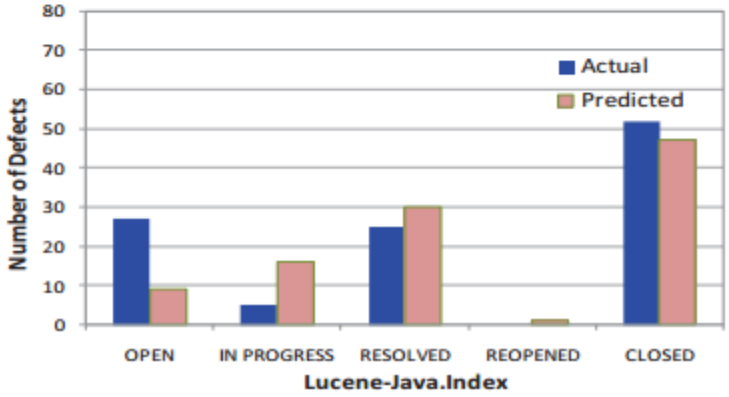












深入理解PageRank

其他应用

1. 网站导航能力评价

Y. Zhou, H. Leung, P. Winoto. MNav: A Markov model based web site navigability measure. IEEE TSE, 2007

2. Aspect自动挖掘

C. Zhang, H. Jacobsen. Mining crosscutting concerns through random walks. IEEE TSE, 2012

3. 推荐系统

Z. Saul, V. Filkov, P. Devanbu. Recommending random walks. FSE 2007

4. 宏观经济分析

刘振亚等. 解密复兴科技-基于隐蔽马尔科夫模型的时序分析方法, 2014

5. 语音信号识别

B. Juang, L. Rabiner. Hidden Markov models for speech recognition. Technometrics 1991

Thanks for your time and attention!