基于 EM-GA 改进贝叶斯网络的研究及应用 *

金俊丽[1], 赵 川[2], 杨 洁[3]
(1. 淮海工学院 理学院, 江苏 连云港 222005; 2. 重庆大学 机械工程学院, 重庆 400030; 3. 重庆通信学院, 重庆 400035)

摘 要: 为了解决软件风险分析中可能出现的数据不完整以及影响因素间关系复杂的问题,提出了一种改进贝叶斯网络的软件项目风险分析方法。将遗传算法和 EM 算法相结合得到 EM-GA 算法,利用 EM-GA 算法对软件项目分析过程中贝叶斯网络结构中的参数进行学习,同时优化网络结构,通过实例验证了该方法的有效性及可行性。

关键词: 贝叶斯网络; EM-GA 算法; 软件项目; 风险分析
中图分类号: TP311.52 **文献标志码**: A **文章编号**: 1001-3695(2010)04-1360-03
doi:10.3969/j.issn.1001-3695.2010.04.041

Research on Bayesian network improved by EM-GA and its application

JIN Jun-li[1], ZHAO Chuan[2], YANG Jie[3]
(1. *College of Science, Huaihai Institute of Technology, Lianyungang Jiangsu* 222005, *China*; 2. *College of Mechanical Engineering, Chongqing University, Chongqing* 400030, *China*; 3. *Chongqing Communication College, Chongqing* 400035, *China*)

Abstract: In order to solve the problem of incomplete data and complex relations among influencing factors which may appear in the software risk analysis, this paper presented a software project risk analysis process based on Bayesian networks which has been improved. Firstly, presented a EM-GA algorithm based on genetic algorithm. Then, used the algorithm to optimize the Bayesian networks structures and solve Bayesian parameter learning. Finally, the experiment results show this algorithm provide a new method for software project risk analysis process.

Key words: Bayesian networks; EM-GA algorithm; software project; risk analysis

0 引言

当前软件开发规模大幅增加、软件开发环境复杂化以及软件运行环境急剧恶化时,软件开发项目面临着更加严峻的挑战。在软件开发阶段,仅考虑通过技术手段来提高软件质量变得更加困难。许多研究者意识到,保证软件质量的基本问题不仅在于是否使用新技术,更在于该过程中是否有科学的风险管理方法和流程[1],加强软件开发过程中风险管理以确保软件开发质量是当前软件工程研究领域中亟待解决的重要问题。围绕这一问题,国外近年来出现了 CMMI、SERIM、SRAM 模型以及贝叶斯网络方法等软件过程风险管理及评价技术,用于软件开发过程风险管理领域[2]。然而,面向软件开发过程的风险管理是一个涉及人、技术、组织、软件产品和环境的复杂过程。环境、需求、技术及信息的不断演化造成了软件开发风险管理过程的动态性、不确定性与复杂性。所以,已有的模型和方法较难实现软件开发过程风险因素的快速有效的识别、分析与评价。有鉴于此,本文提出基于 EM-GA 改进贝叶斯网络的软件项目风险评价方法。

软件风险管理可以分为五个步骤,即风险识别、风险分析、风险计划、风险跟踪和风险控制,如图 1 所示。风险评估包括风险识别与风险分析,是软件风险管理中非常重要的一环。对软件项目进行风险评估研究能够帮助项目管理人员很好地开展风险控制计划,有效地实施项目管理,因此对于项目开发过程中可能遇到的风险必须作出合理的评价[1,3,4]。

本文提出一种基于改进贝叶斯网络的软件项目风险分析方法。首先,介绍了贝叶斯网络及遗传算法的理论基础;然后,将遗传算法应用于贝叶斯网络结构学习,建立了软件项目风险分析的 EM-GA 算法;最后,采用实例对该方法进行了验证,证明该模型可以对项目风险作出较为客观的分析。

1 基础概念

1.1 贝叶斯网络

贝叶斯网络是一种概率推理技术,用来表示变量间概率依赖关系的图形模式,它以概率理论为数学基础,处理在描述不同知识成分之间的条件相关而产生的不确定性,提供了一种将知识直观地图解可视化的方法[5,6]。

贝叶斯网络由两部分组成:贝叶斯网络结构和贝叶斯条件概率表。给定一个随机变量集 $\chi = \{X_1, X_2, \cdots, X_n\}$, χ 是一个 m 维向量。贝叶斯网络说明 χ 上的联合条件概率分布。贝叶斯网络定义如下:

$$B = \langle G, \theta \rangle \tag{1}$$

G 是一个有向无环图,其顶点对应于有限集 χ 中的随机变量 $X_1, X_2, \cdots, X_n$。其弧代表一个函数依赖关系。如果有一条

收稿日期: 20009-07-05; **修回日期**: 2009-09-03 **基金项目**: 国家教育部"新世纪优秀人才支持计划"资助项目(NCET-07-0908)
作者简介:金俊丽(1973-),女,吉林洮南人,讲师,硕士研究生,主要研究方向为智能算法、风险管理;赵川 (1984-),男,河北石家庄人,硕士研究生,主要研究方向为风险管理(zhaochuan@cqu.edu.cn);杨洁(1973-),女,重庆人,讲师,博士,主要研究方向为创新管理、风险管理.

弧由变量 Y 到 X，则 Y 是 X 的双亲或者直接前驱，而 X 则是 Y 的后继。一旦给定其双亲，图中的每个变量独立于图中该节点的非后继。在图 G 中 X_i 的所有双亲变量用集合 $pa(X_i)$ 表示。

θ 代表用于量化网络的一组参数。对于每一个 X_i，每个 $pa(X_i)$ 的取值 x_i 存在如下一个参数：

$$\theta_{x_i|pa(x_i)} = P(x_i|pa(X_i)) \tag{2}$$

它指明了在给定 $pa(X_i)$ 条件下 x_i 事件发生的条件概率。因此实际上一个贝叶斯网络给定了变量集合 χ 上的联合条件概率分布：

$$P_B(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} P_B(X_i|pa(X_i)) \tag{3}$$

1.2　遗传算法

遗传算法(genetic algorithm)是模拟自然界生物群体进化过程的一种随机优化方法，具有不依赖于问题模型的特征、寻优过程的自适应、隐含的并行性、解决复杂的非线性问题的鲁棒性以及对优化目标函数无须连续、可微等苛刻条件等优点，并在许多复杂优化问题的应用中都得到令人满意的解。

在遗传算法中，将 n 维决策向量 X 用 n 个记号 $X_i(i=1,2,\cdots,n)$ 所组成的符号串表示为 $X = X_1X_2\cdots X_n$。其中每个 X_i 可以看成一个遗传基因，它的所有可能取值称为等位基因。这样 X 可以看成是由 n 个遗传基因所组成的染色体。一般情况下，染色体的长度 n 是固定的，但对某些问题 n 也可以是变化的。编码所组成的排列形式是个体 X 的基因型，与之对应的值是个体 X 的表现型。通常个体的表现型与基因型是一一对应的，但有时也允许基因型与表现型是多对一的关系。对于每一个个体 X，要按照一定的规则确定其适应度。个体 X 的适应度与其对应的表现型的目标函数值相关联，X 越接近目标函数的最优点，其适应度越大；反之，其适应度越小[7,8]。

生物的遗传过程主要是通过染色体之间的重组和染色体变异来完成的。与此对应，遗传算法中最优解的搜索过程也模仿生物的遗传过程，使用遗传算子作用于群体 $P(t)$，得到下一代群体 $P(t+1)$。

1.3　*EM* 算法

给定观测到的数据 $D = \{D_1, D_2, \cdots, D_n\}$，未观测到的数据 $Z = \{Z_1, Z_2, \cdots, Z_n\}$，参数概率分布 $P(Y|\theta)$。其中：Y 为全部训练样本，$Y = \{Y_1, Y_2, \cdots, Y_n\}$，$Y_i = D_i \cup Z_i$，$\theta$ 为参数。

结果：求参数 θ，使得期望 $E[\ln p(Y|\theta)]$ 最大[9]。

使用观测数据集 $D = \{D_1, D_2, \cdots, D_n\}$ 和当前参数 θ，计算 $Y = D \cup Z$ 的自然函数期望 $Q(\theta^{(l)}|\theta)$：$Q(\theta^{(l)}|\theta) \leftarrow E[\ln p(Y|\theta^{(l)})|\theta, D]$。从当前的估计值 θ 到下一个估计值 $\theta^{(l)}$ 需要两个步骤：期望计算(E-Step)和最大化计算(M-step)。

1)期望计算　它是计算给定可观测训练样本集 D 和当前 θ 时，数据全集 Y 的概率分布期望：

$$Q(\theta^{(l)}|\theta) = E[\ln p(Y|\theta^{(l)})|\theta, D] = \sum_{l}\sum_{Z_i} \ln p(D_l, Z_l|\theta) p(Z_l|D_l, \theta^l) \tag{4}$$

2)最大化计算　它是最大化当前函数 $Q(\theta^{(l)}|\theta)$，具体方法是选择参数值 $\theta^{(l)}$ 代替 θ，使得函数 $Q(\theta^{(l)}|\theta)$ 最大：

$$\theta^{(l)} = \max Q(\theta^{(l)}|\theta) \tag{5}$$

2　软件项目风险分析的 EM-GA 算法

在一个软件项目开发中，由于影响因素众多，而且因素间关系相当复杂，会导致对于系统结构知之甚少，因素间的依赖关系也不知道。这时就需要先进行结构学习，再进行参数学习。对此，本文提出了遗传算法与 EM 算法相结合的 EM-GA 算法，EM 算法用于优化网络参数，遗传算法用于结构学习。算法的基本思想是：首先从初始群体中任意选择一个个体，也就是一个初始的网络结构；根据这个网络结构与不完备数据计算 EM 算法的 E 步(求期望)的计算；然后对 E 步的结果求极大值，得到补充的数据；在补充完整的数据之上进行遗传算法操作；一轮完成后，得到一个最优个体，在这个最优个体上，重复 EM 算法。

2.1　编码方式

由贝叶斯网络的定义可知其网络结构是有向无环图，所以可以将网络中的每一个变量的编码用其父节点的集合来表示。整个网络的编码就是按一定顺序排列的节点的父节点集合，如对图 2 中的四个节点组成的有向无环图的编码为 $[X_1|;X_2|X_1;X_3|X_1;X_4|X_2,X_3]$，用分号来把各父节点集分开，“|”前边是要表示的节点，“|”后边是该节点的父节点集。每个节点的局部结构(该节点和其父节点集)都可以表示成位置的一个分量形式。

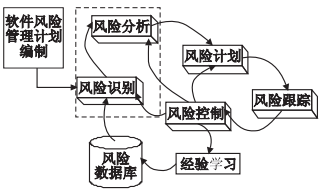

图1　软件风险管理框架图

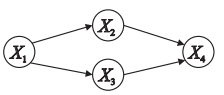

图2　一个简单的有向无环图

2.2　算法过程描述

算法过程如下：

a)使用种群中最优个体(如果是算法首次执行，则随机产生一个初始网络)S_c，利用 EM 算法对不完备数据集 D 进行完备化，得到完整数据集 D_c。

b)如果算法是首次执行，那么就随机生成初始化网络结构。否则直接执行 c)，进行下一轮进化。

c)对初始群体 S_Δ，按照概率 P_c 或 P_m 进行交叉或变异操作，如果变异在该种群最优秀的个体 t 上发生，那么必须在变异前和变异的两个个体中选择更坏的个体，而其余变异操作后不作选择。这样就得到变异后群体 S_Δ'。

d)对于 S_Δ' 中的每一个网络 S，进行如下几步操作：

(a)如果网络 S 不满足每个变量最多有 m 个父节点的约定，则采用局部优化方法，选出不超过 m 个父节点集合。

(b)检查网络 S 是否包含非法结构，及判断 S 中是否存在有向环。如果网络 S 包含非法结构，则给网络 S 赋予一个较小的适应度；否则，按照下面的公式计算其适应度 Fs：

$$F_s = \text{MDL\_score}(S \times D_c) \tag{6}$$

其中：MDL_score($S \times D_c$)就是网络结构 S 与数据集 D_c 的 DML 评价。

c)计算网络结构 S 的被选择概率 P_s：

$$P_s = \text{rank}(F_s)/[r_j(r_j+1)/2] \tag{7}$$

其中：rank(F_s)表示 F_s 的等级，它的值根据 F_s 的大小来划分；r_j 表示进化群体的规模。

e)按照 S_Δ' 中每个网络 S 的选择概率 P_s 和轮盘赌的机制选取 r_j 个个体进入下一代，并更新 S_Δ'；然后从更新的种群中选

择最优个体 S'。如果 S'优于迄今为止的最佳个体 S,则当前最差个体被迄今为止的最佳个体取代,并复制当前的最佳个体 S'作为迄今为止的最佳个体 S。

f)选择最佳个体 S',使得 $F_{s'} = \max\arg(F_s)$。如果 $F_{s'} > F_{sc}$,则 $S_c = S'$。

判断算法终止条件是否满足,如果满足则退出,否则转 a)继续进化。

3　实证分析

选取软件项目开发中的软件使用者(D_1)、项目复杂程度(D_2)、项目战略定位(D_3)、项目范围(D_4)、资源安排(D_5)、软件需求(D_6)、组织环境(D_7)、开发团队(D_8)、计划与控制(D_9)九项因素。表 1 的学习样本为 100 个软件项目的实际数据,该数据已经过离散化处理。

具体步骤如下:

a)用计算机程序创造一些较好的网络结构作为初始群体,当前的候选网络从初始群体中随机产生;

b)利用 EM 算法对不完备数据集进行完备化,得到完整数据集;

c)由遗传算法进行优化,选择方式为轮盘赌法,进化代数为 100 代,交叉概率为 0.05,交叉方式为均匀交叉,变异概率为 0.05。

经过 100 代的进化,得到最优解,其对应的最佳网络结构如图 3 所示。

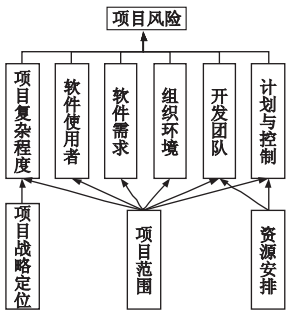

图3　最佳网络结构图

本文从算法计算精确度及算法收敛性这两个方面验证该参数学习的有效性。图 4 和 5 为项目复杂程度节点的进化过程,可以看出,本节点在进化过程中没有退化现象。

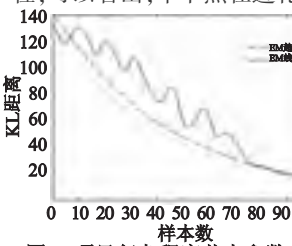

图4　项目复杂程度节点参数学习的KL图

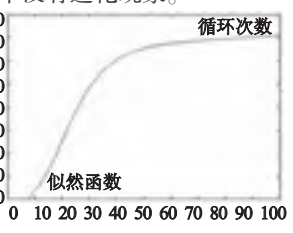

图5　项目复杂程度节点循环次数与似然函数趋紧关系

图 4 为项目复杂程度节点参数学习的 KL 距离,当 KL 距离越小时,参数学习的结果越精确;图 5 为项目复杂程度节点循环次数与似然函数的趋紧关系,算法大概经过 70 个循环就估计出最大函数,因此算法具有较好的收敛性。

将表 2 的实验样本代入已经构建的贝叶斯网络,经过运算后得到的风险率为[0.761 0.350 0.275 0.492 0.339],风险率误差绝对值小于 0.001。由此证明,该软件项目风险分析模型是有效的。

表 1　学习样本集

序号	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	风险率
1	3	2	2	3	3	2	2	3	4	0.399
2	3	4	3	5	4	3	4	3	4	0.114
3	1	1	2	2	1	1	1	2	1	0.646
4	2	3	3	4	3	2	3	3	2	0.264
5	3	1	2	3	3	1	1	2	2	0.491
ú	ú	ú	ú	ú	ú	ú	ú	ú	ú	ú
100	2	3	3	3	3	4	3	2	2	0.308

表 2　实验样本集

序号	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	风险率
1	1	2	2	3	1	2	2	2	3	0.762
2	3	2	3	2	4	2	4	2	3	0.349
3	1	3	2	5	1	4	1	3	2	0.276
4	2	2	3	2	3	5	3	1	4	0.491
5	2	5	3	4	2	1	1	3	3	0.338

4　结束语

本文提出一种新的基于改进贝叶斯网络的软件项目风险分析方法。将遗传算法与 EM 算法相结合,提出了 EM-GA 算法,该算法可以用来构建风险分析的贝叶斯网络结构与参数学习;采用实例对算法进行了验证,经过遗传优化以后,得到了最佳软件项目风险分析贝叶斯网络结构,而且其参数精度也达到预定要求。基于 EM-GA 算法软件项目风险分析模型与传统软件风险分析模型相比:模型结构构建更加合理,参数更加精确,有效减少了人为构建模型和参数选择带来的主观性,使软件风险分析更加科学、合理。

参考文献:

[1] BOEHM B W. Software risk management: principles and practices [J]. IEEE Sofeware, 2007,8(1):32-41.

[2] LIU Xiao-qing, KANE G, BAMBROO M. An intelligent early warning system for software quality improvement and project management[C]// Proc of the 15th IEEE International Conference on Tools with Artificial Intelligence. Washington DC:IEEE Computer Society, 2003:32.

[3] KLASCHKE G. What the CHAOS chronicles 2003 reveal[R]. San Diego: Cost Xpert Group, 2004.

[4] AKHTE A. Requirement reliability metrics for risk assessment[C]// Proc of Student Conference on Engineering Sciences and Technology. [S.l.]:NED University of Engineering, 2004: 452-461.

[5] CHARETTE R. Software engineering risk analysis and management [M]. New York: McGraw Hill, 2006: 178-189.

[6] HOUSTON D X, MACKULAK G T, COLLOFELLO J S. Stochastic simulation of risk factor potential effects for software development risk management [J]. The Journal of Systems and Software, 2001, 59(3):247-257.

[7] 冯楠,李敏强,寇纪淞,等.基于贝叶斯网络的软件项目风险分析过程[J].计算机工程与应用,2006,42(18):16-18.

[8] 黄友平.贝叶斯网络研究[D].北京:中国科学院研究生院,2005.

[9] 黄浩,宋瀚涛,陆玉昌.基于小生境遗传算法的贝叶斯网络结构学习算法研究[J].计算机应用研究,2007,24(4):100-103.

[10] 葛继科,邱玉辉,吴春明,等.遗传算法研究综述[J].计算机应用研究,2008,25(10):2911-2916.

[11] 张少中,杨南海,王秀坤.贝叶斯网络参数的在线学习算法及应用[J].小型微型计算机系统,2004,25(10):1800-1801.