

基于遗传算法的软件项目风险辨识研究

冯楠¹,李敏强¹,方德英²

FENG Nan¹,LI Min-qiang¹,FANG De-ying²

1.天津大学 管理学院,天津 300072

2.北京联合大学 商务学院,北京 100025

1.School of Management,Tianjin University,Tianjin 300072,China

2.Business College,Beijing Union University,Beijing 100025,China

E-mail:fengnan_1978@yahoo.com.cn

FENG Nan,LI Min-qiang,FANG De-ying.Research on software project risk identification based on Genetic Algorithms. Computer Engineering and Applications,2008,44(9):7-9.

Abstract: A software project risk identification process based on Genetic Algorithms(GA) is presented in this paper.The risk identification process utilizes GA to search the rules of risk identification based on historical data in software project in order to identify the risks in the process of software project development.The application process of GA in risk identification is analyzed in the paper.Finally,the validity of risk identification process is proved using an example.

Key words: software project;risk identification;Genetic Algorithms(GA)

摘 要: 该文提出了一种基于遗传算法的软件项目风险辨识过程。该风险辨识过程以软件项目历史数据为基础利用遗传算法搜索风险辨识规则,从而达到在软件开发过程中识别风险的目的。深入分析了遗传算法在风险辨识中的应用过程。最后,通过具体实例证明该风险辨识过程的有效性。

关键词: 软件项目;风险辨识;遗传算法

文章编号:1002-8331(2008)09-0007-03 **文献标识码:**A **中图分类号:**TP311.52;O212

1 引言

如图1所示,软件项目风险管理过程包括风险辨识、风险分析、风险计划、风险跟踪以及风险应对这5个基本过程元素^[1]。

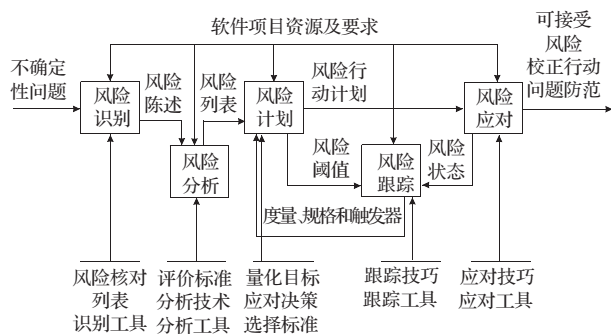


图1 软件项目风险管理过程

而软件项目风险管理过程当中风险辨识是最基础的环节,其它风险管理过程的进行都必须以风险辨识的结果为基础。风险辨识所采用的方法和工具主要有决策树、故障树、风险源清单、风险档案表以及阶段风险报告和等等。但这些方法大多以专家的主观经验为基础,并且不适合规模化操作。而利用

遗传算法进行风险辨识是基于软件项目历史数据搜索风险辨识规则,从而达到在软件开发过程中识别风险的目的。同其它辨识方法相比,由于遗传算法简单明了,适合于大规模并行计算,因此能够更有效地应用于类似软件项目的复杂系统。同时,遗传算法是从一群初始点开始搜索,而不是从单一初始点开始搜索。因此,得到问题全局最优解的概率大大提高了。

2 软件项目风险辨识

所谓风险辨识就是采取有严格计划的步骤,在妨碍项目成功的因素变成问题之前发现并定位它们。风险辨识是软件项目风险管理过程的第一步。

风险辨识过程定义描述如图2所示。

风险辨识模块定义描述了从输入变为输出的过程。该过程包括输入、输出、约束条件以及保障机制4个过程。

(1) 输入

不确定性问题是指未知的事物,是被假定和怀疑的一部分。而将历史数据加入到风险辨识过程当中有利于增强风险辨识的准确性。

(2) 输出

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70371046);北京市拔尖创新人才基金资助。

作者简介:冯楠(1978-),男,博士,研究方向:软件工程,信息管理与信息系统;李敏强(1965-),男,博士,教授,博士生导师,研究方向:系统工程与信息系统,人工智能;方德英(1964-),男,博士,教授,硕士生导师,研究方向:信息系统和系统工程。

收稿日期:2007-08-07

修回日期:2007-12-03

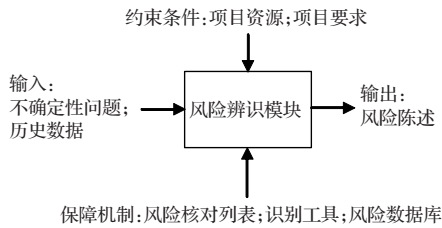


图2 风险辨识过程

风险陈述是用简单的表示法对风险进行简要说明。这一表示法为:问题、可能性和结果。风险陈述可分解为主题和风险的两个主要属性——概率和结果。风险分析模块根据风险陈述对风险进一步分析。

(3) 约束条件

软件项目资源用成本、时间和人员来限制风险辨识的范围。成本有限时,可以采用相对成本较低的方法来识别风险;项目周期较短时,可以采取更快的识别风险的方法。当然,如果因为项目资源不足而采取了缩减措施后,就会有降低风险辨识效果的风险。

(4) 保障机制

风险核对列表包括与风险核对主题相关的典型风险区域。风险核对列表能通过各种形式组织风险,如合同类型、成熟度级别、生命周期模型、开发阶段、组织结构、项目规模和应用技术等。它们有助于在指定区域里完成风险辨识。

风险数据库是一个已知风险和相关信息的仓库,它将风险输入计算机,并分配下一个连续的号码给这个风险,同时维持所有已识别风险的历史记录。

3 遗传算法在风险辨识中的应用过程

遗传算法抽象于生物体的进化过程,通过全面模拟自然选择和遗传机制,形成一种具有“生成+检验”特征的搜索算法。遗传算法以编码空间代替问题的参数空间,以适应度函数为评价依据,以编码群体为进化基础,以对群体中个体位串的遗传操作实现选择和遗传机制,建立起一个迭代过程。在这一过程中,通过随机重组编码位串中重要的基因,使新一代的位串集合优于老一代的位串集合,群体的个体不断进化逐渐接近最优解,最终达到求解问题的目的。

风险辨识过程中,风险规则可以表示为:

$$R_j: \text{IF } rf_{1j} \ \& \ rf_{2j} \ \& \ \dots \ \& \ rf_{nj} \ \text{ THEN } \omega$$

其中, $rf_{ij} (i=1, 2, \dots, n)$ 表示风险规则前提的属性值, ω 表示风险规则的结论。“属性”对应风险辨识过程中的风险因素;“结论”对应风险辨识过程中的风险。

基于遗传算法的软件项目风险辨识过程包括以下几个步骤:

(1) 编码

在利用遗传算法搜索风险规则空间、求取最优风险规则之前,首先要进行编码。对于分类问题的产生式规则而言,二进制编码是一种简单直观的编码方式。将风险规则前提用一个二进制字符串表示,若某位为1表示具有该属性,为0则表示无此属性。由0,1组成的有序字符串表示用于识别某一分类类别 ω 的属性组合。编码后利用遗传算法不断调整组合方式,使其性能逐渐达到最优。

(2) 群体规模

随机产生 n 个二进制编码链,每个编码链代表遗传算法的原始群体。

(3) 确定适应性评价函数

适应值函数 $F(R_j)$ 由相似度函数 $F_1(R_j)$ 和信息测度函数 $F_2(R_j)$ 两部分组成。

风险规则 R_j 与历史数据中某正例 E_j 、反例 C_k 的相似度分别为:

$$\varphi\langle R_j, E_j \rangle = \frac{\sum_{i=1}^n rf_{ij} e_{ij}}{\sqrt{\sum_{i=1}^n (rf_{ij})^2} \sqrt{\sum_{i=1}^n (e_{ij})^2}} \quad (1)$$

$$\varphi\langle R_j, C_k \rangle = \frac{\sum_{i=1}^n rf_{ij} c_{ik}}{\sqrt{\sum_{i=1}^n (rf_{ij})^2} \sqrt{\sum_{i=1}^n (c_{ik})^2}} \quad (2)$$

若数据库中共有 l 个正例 m 个反例,则风险规则 R_j 与历史数据中实例之间的相似度函数 $F_1(R_j)$ 为:

$$F_1(R_j) = \frac{\sum_{j=1}^l \varphi\langle R_j, E_j \rangle}{\left(1 + \frac{1}{m} \sum_{k=1}^m \varphi\langle R_j, C_k \rangle\right) \cdot l} \quad (3)$$

$F_1(R_j)$ 越大,表示风险规则 R_j 与正例集中的正例越相似,而与反例集中的反例越不相似,风险规则 R_j 也就越合适。

设两个离散随机变量 X 和 Y , 变量 Y 取值 y 的先验概率为 $p(y)$, 而变量 X 取值 x 后,变量 Y 取值 y 的后验概率为 $p(y|x)$ 。在信息熵的概念中,为描述 IF $X=x$ THEN $Y=y(p(y|x))$ 的信息含量,有如下的测度函数:

$$J(Y; X=x) = \sum_y p(y|x) \log(p(y|x)/p(y)) \quad (4)$$

对于二值离散变量,上式可转化为:

$$J(Y; X=x) = p(y|x) \log(p(y|x)/p(y)) + (1-p(y|x)) \log((1-p(y|x))/(1-p(y))) \quad (5)$$

在风险规则优化过程中,不仅要求风险规则的信息量大,而且要求规则比较简单。即 R_j 的前提 rf_{ij} 出现的概率越大越好。为此,在信息测度函数中引入概率因子 $p(x)$:

$$F_2(R_j) = p(x) \cdot J(Y; X=x) \quad (6)$$

则综合适应值评价函数 $F(R_j)$ 为:

$$F(R_j) = F_1(R_j) + F_2(R_j) \quad (7)$$

(4) 选择

利用轮盘赌方式实现适应值比例选择方法。即每个个体被选择的期望数量与其适应值和群体平均适应值的比例有关。

对于给定的规模为 n 的群体 $P = \{a_1, a_2, \dots, a_n\}$, 个体 $a_j \in P$ 的适应值为 $f(a_j)$, 其选择概率为:

$$p_s(a_j) = \frac{f(a_j)}{\sum_{i=1}^n f(a_i)}, \quad j=1, 2, \dots, n \quad (8)$$

该式决定后代种群中个体的概率分布。经选择操作生成用于繁殖的交配池,其中父代种群中个体生存的期望数目为:

$$p(a_j) = n \cdot p_s(a_j), \quad j=1, 2, \dots, n \quad (9)$$

(5) 交叉

首先从交配池中随机取出要交配的一对个体;然后根据位串长度 L , 对要交配的一对个体随机选取 $[1, L-1]$ 中一个或多个

的整数 k 作为交叉位置;最后根据交叉概率 $p_c(0 < p_c \leq 1)$ 实施交叉操作,配对个体在交叉位置处相互交换各自的部分内容,从而形成新的一对个体。

(6) 变异

变异算子通过按照变异概率 $p_m(0 < p_m \leq 1)$ 随机反转某位等位基因的二进制字符值来实现。对于给定的染色体位串 $s = a_1, a_2, \dots, a_l$, 变异过程如下:

$$O(p_m, x): a'_i = \begin{cases} 1 - a_i, & x_i \leq p_m \\ a_i, & \text{otherwise} \end{cases} \quad i \in \{1, 2, \dots, L\} \quad (10)$$

生成新的个体 $s' = a'_1, a'_2, \dots, a'_l$ 。其中 x_i 是对应于每一个基因位产生的均匀随机变量, $x_i \in [0, 1]$ 。

(7) 迭代过程

由产生的新风险规则替代原有的风险规则,然后转步骤 3 继续优化。最后,在采用精英保留选择策略的情况下,按每代最佳个体的适应值变化情况确定迭代终止条件。

(8) 规则解码

进化结束后,将以编码形式表示的规则解码成形如式(1)的风险规则形式。

4 实证分析

本文以进度延迟、高错误率和成本增加这三个软件项目风险为例说明该风险辨识过程的有效性。表 1 和表 2 为从风险数据库中提取出来的 3 类风险的辨别实例以及相关的 20 种风险因素。以该数据为基础,利用遗传算法搜索辨别这 3 类风险的最优风险规则。

表 1 软件项目风险因素

| 序号 | 风险因素 | 序号 | 风险因素 |
|----|--------|----|----------|
| 1 | 缺少关键技术 | 11 | 过量文档工作 |
| 2 | 依赖关键人物 | 12 | 反复修改需求分析 |
| 3 | 人员流动性大 | 13 | 项目规模复杂 |
| 4 | 缺少责任感 | 14 | 技术不成熟 |
| 5 | 低士气 | 15 | 外部接口复杂 |
| 6 | 低效率 | 16 | 缺乏项目管理 |
| 7 | 缺少客户支持 | 17 | 开发者缺乏交流 |
| 8 | 错误度量标准 | 18 | 进度压力大 |
| 9 | 缺少历史数据 | 19 | 错误预算估计 |
| 10 | 配置控制不足 | 20 | 项目缺乏经验 |

在风险辨识过程中,设群体规模为 $n=25$,交叉概率 $p_c=0.8$,变异概率 $p_m=0.05$,迭代次数 $T=60$ 。如表 3 所示,对每一类风险,从搜索到的规则集合当中取性能最优的作为该风险的辨识规则。

根据表 3 可以生成相应的风险辨识规则:

(上接 6 页)

- [28] Park B U, Sickles R C, Simar L. Stochastic panel frontiers: A semiparametric approach[J]. Journal of Econometrics, 1998, 84(2): 273-301.
- [29] 江亦曼, 乔晓春, 于学军, 等. 中国未来人口发展与生育政策研究[J]. 人口研究, 2000, 24(3): 18-34.
- [30] 刘介明, 石庆泉, 孙洪亮. 医疗事故风险分析及人工神经网络预测模型[J]. 武汉理工大学学报, 2004, 26(1): 70-74.
- [31] 外科手术的风险预测与临床评价[J]. 临床外科杂志, 2006, 14(9): 540-541.
- [32] 张润森, 陈绍军. 最低生活保障资金供求的系统模型[J]. 统计与决

表 2 软件项目风险辨别实例

| 序号 | 风险类别 | 风险实例 |
|----|------|--|
| 1 | 进度延迟 | (1, 6, 18)(1, 2, 18)(1, 5, 14)(1, 2, 6, 18) |
| | | (2, 4, 17, 18)(6, 18)(12, 14, 18)(1, 2, 6, 12) |
| | | (5, 6, 12, 14, 18)(5, 7, 18, 20)(1, 6, 14, 18) |
| 2 | 高错误率 | (1, 5, 18)(13, 18)(3, 7, 13, 17, 18)(3, 4, 1, 3, 18) |
| | | (10, 13, 14)(3, 14)(3, 5, 7, 14)(11, 14, 17, 18) |
| 3 | 成本增加 | (1, 3, 7, 14)(3, 10, 11, 14)(3, 13, 14) |
| | | (1, 6, 12)(12, 18, 19)(1, 6, 19)(1, 5, 14, 19) |
| | | (1, 12, 19)(1, 6)(1, 18, 19)(6, 19, 20) |

表 3 搜索结果

| 序号 | 属性值 | 结论 |
|----|------------------------|------|
| 1 | (10000100000000000100) | 进度延迟 |
| 2 | (00100000000011000100) | 高错误率 |
| 3 | (10000100000000000010) | 成本增加 |

规则 1 IF 缺少关键技术 & 低效率 & 进度压力大 THEN 进度延迟。

规则 2 IF 人员流动性大 & 项目规模复杂 & 技术不成熟 & 进度压力大 THEN 高错误率。

规则 3 IF 缺少关键技术 & 低效率 & 错误预算估计 THEN 成本增加。

5 结束语

软件项目风险管理过程当中风险辨识是最基础的环节。

参考文献:

- [1] Friedman N.A. Bayesian approach to structure discovery in Bayesian networks[J]. Machine Learning, 2003: 95-125.
- [2] Riggelsen, Carsten. Learning parameters of Bayesian networks from incomplete data via importance sampling[J]. International Journal of Approximate Reasoning, 2006, 42(1): 69-83.
- [3] 冯楠, 李敏强, 寇纪淞. 一种改进的软件项目投资风险评价模型[J]. 计算机工程, 2007, 33(7): 41-44.
- [4] 张少中. 基于贝叶斯网络的知识发现与决策应用研究[D]. 大连: 大连理工大学, 2003: 74-76.
- [5] Madsen A L, Lang M, Kjaerulff U B. The Hugin tool for learning Bayesian networks[C]// Proc of Symbolic and Quantitative Approaches to Reasoning with Uncertainty of 7th European Conference. Berlin, Germany: Springer-Verlag, 2003: 594-605.
- [6] Wang Shuang-Cheng, Yuan Sen-Miao. Research on learning Bayesian networks structure with missing data[J]. Journal of Software, 2004, 14(7): 1042-1048.
- [7] 策, 2005(20): 56-58.
- [33] Pay D, Lucy G, Dagmar K, et al. Population density, soil nutrient depletion, and economic growth in sub-Saharan Africa[J]. Ecological Economics, 2001, 38: 251-258.
- [34] Morris J N, Deeming C. Minimum incomes for healthy living (MIHL): next thrust in UK social policy? [J]. Policy & Politics, 2004, 32(4): 441-454.
- [35] 欧世平. 我国农村养老保障体系建设研究[D]. 合肥: 安徽大学, 2006.
- [36] Immervoll H, Sutherland H, De Vos K. Child poverty and child benefits in the European Union[R]. Euromod Working Papers, em/00, 2000: 1-31.