粗糙集和贝叶斯网络在软件风险评估中的应用①

任雪利 (曲靖师范学院 计算机科学与工程学院 云南 曲靖 655011)

摘　要： 准确的风险评估为项目的顺利进行提供了理论依据，采用贝叶斯网络对软件开发过程的风险进行定量的评估，为了降低无关属性对评估结果的影响，引入了粗糙集理论。

关键词： 粗糙集；风险评估；贝叶斯网络

Application of Bayesian Network and Rough Sets in Software Risk Estimation

REN Xue-Li

(Department of Computer Science and Engineering, Qujing Normal University, Qujing 655011, China)

Abstract: The paper proposes a solution to risk estimation and uses Bayesian network and Rough Set to estimate the subjection of prior probability.

Keywords: risk estimation; bayesian network; rough sets

软件项目管理是使软件项目能够按照预定的成本、进度、质量顺利完成，在评定和计划一个项目时，必须关注可能使项目偏离原定目标的风险。风险管理包括风险评估与风险控制 2 个部分； 其中：风险评估包括风险识别、风险分析、风险计划，风险控制包括风险跟踪和风险应对。风险识别是风险管理过程的第 1 步,主要是识别风险和风险来源，将项目的不确定性因素及问题转换为具体的可以被描述和估量的风险，在影响到项目之前揭示出主要风险。在风险识别完之后,用定性或定量的工具对风险列表进行分析，对风险列表中的各项风险进行排序，将风险估计量或数据转化为一种用来确定优先决策的形式,为项目管理者专注于正确的、重要的风险提供行动基础[1]。贝叶斯网络不仅具有很强的推理能力，而且更能反映容易理解的推理过程[2]。准确的推理依赖于合理的贝叶斯结构的建立，如果贝叶斯网络中加入了对结果没有任何影响的属性节点，不仅增加了网络的复杂性，而且会导致概率估计的不准确，为了减少无关属性对评估结果的影响，本文采用粗糙集对属性进行约简，然后根据节点间的因果关系建立贝叶斯网络来进行风险评估。

1　贝叶斯网络与粗糙集

1.1 贝叶斯网络

贝叶斯网络是一种基于概率推理的图形化网络，结点用随机变量标识，图中结点 V_i 条件独立于 V_i 的父结点给定的 V_i 的非后代结点构成的任何结点子集[3,4]。即假设 $A(V_i)$是图中非 V_i 后代结点的任何结点集合，设 $P(V_i)$是图中 V_i 的直接双亲，贝叶斯网络图中所有 V_i， $P(V_i \ A(V_i)P(V_i))=P(V_i|P(V_i))$。定义信念为 $Bel(x)=P(x|e)$，即在有已知证据 e 的情况下，事件 x 发生的条件概率，反映在一定环境下某一事件发生的可能性。e 表示为 $e=e^-x \cup e^+x$ ，其中 e^-x 反映以 x 为根结点的子树，e^+x 反映树的其余部分，则信念可表示为 $Bel(e)=P(x|e^-x, e^+x)= \alpha *P(e^-x|e^+x,x)* P(x|e^+x)= \alpha *P(e^-x|x)*P(x|e^+x)$，式中 $\alpha =[P(e^-x|e^+x)]^{-1}$ 为归一化因子。令 $\lambda(x)=P(e^-x|x)$,表示对诊断的支持；$\pi(x)=P(x|e^+x)$,表示对预报的支持，则有 $Bel(x)= \alpha * \lambda(x)* \pi(x)$.在实际推理过程中，当证据 e 到来时，网络信念得到更新。

1.2 粗糙集

粗糙集(Rough Set) 理论是波兰科学家 Pawlak 在 1982 年提出的一种用于分析不确定性数据的数学

① 基金项目:曲靖师范学院校级科研项目(2008QN007)
收稿时间:2009-09-01;收到修改稿时间:2009-11-15

理论，该理论已成功应用于数据挖掘、机器学习、模式识别等领域，下面简要介绍粗糙集的基本概念[5,6]。

定义 1. 粗糙集中定义信息系统为一个如下四元组：S=(U,A,V,F)，其中 U=(x1,x2,…,xn)是对象集，即论域；A 是属性集合，A=C∪D，且 C∩D=Φ，其中 C 为条件属性，D 为决策属性；V 为属性 A 的值域；F 是 U×A→V 的映射，它为 U 中各对象的属性指定唯一值。S 又称为决策表。

信息系统中的属性并不是同等重要的。属性约简是指可以找到一个较小的属性集，使得可用描述的对象集合必然可用描述。属性约简是粗糙集理论的核心内容，通过属性约简，可以消除冗余属性，减轻评价工作量，提高评价效率。

令 R 为一族等价关系，$r \in R$，如果

$$U/R = U/R - \{r\}$$

则称 r 为 R 中不必要的；否则，则称 r 为 R 中必要的。

如果每一个 $r \in R$ 都为 R 中必要的，则称 R 为独立的；否则称 R 为依赖的。

定义 2. 设 $P \subseteq R$，如果 P 是独立的，且 $U/P = U/R$，则称 P 为 R 的一个约简，记作 $red(P)$。显然，P 可以有多种约简。P 中所有必要关系组成的集合称为 P 的核，记作 $core(P)$，核与约简有如下关系：

$$core(P) = \bigcap red(P)$$

其中，$red(P)$ 表示 P 的所有约简。

1.3 贝叶斯网络与粗糙集的结合

贝叶斯网络具有很强的推理能力，但无关因素不仅增加了网络的复杂性，而且影响推理结果的准确性，使其优势得不到发挥；粗糙集是一种很好的属性约简方法，通过对影响结果的各种因素的计算，剔除与结果无关的因素。如果对粗糙集约简后的属性建立贝叶斯网络，然后再进行评估与推理，这样，不仅发挥了粗糙集的优势，而且充分发挥贝叶斯网络强大的推理能力，对风险进行准确的评估。

2 粗糙集和贝叶斯网络在软件风险评估中的应用

本文建立在粗糙集属性约简的基础上，粗糙集约简后的属性作为建立贝叶斯网络模型的节点，然后进行风险评估以及预测。具体过程可以依照下面的步骤来进行：

1) 收集与待评估的结果有关的信息，并进行离散化。收集影响待评估的结果此前的记录信息，对各因素及结果根据收集的记录确定标准进行离散化，不同的属性可以使用不同的标准，对于数据密集的因素，选用的离散化标准的范围小一些；对于数据稀疏的因素，选用的离散化标准的范围大一些，以免由于同一标准而影响评估结果；

2) 根据各因素对结果的分类情况压缩原信息，剔除对结果没有影响的因素。使用各种因素对所有的记录进行分类，构造原记录的压缩简表，分析该表，剔除对分类结果没有影响的因素；

3) 建立风险预测模型。根据筛选出的因素与结果间的相互影响关系，建立风险的贝叶斯网络模型及节点的先验概率值，作为风险分析及预测的基础；

4) 计算节点的风险概率值。根据建立的模型及节点的先验概率值，进行贝叶斯计算，得到各节点风险发生的概率值，作为制定初始开发计划的依据；

5) 贝叶斯网络信念更新。当获得某一确定的信息时，对网络信念进行更新，根据更新前后网络中各节点概率的比较，可以预测风险是增加了还是减少了，分析导致这一变化的原因，及时调整开发计划并采取有效的措施降低风险发生的概率。

3 粗糙集和贝叶斯网络在风险评估中的应用实例

由于使用贝叶斯网络进行风险评估的具体过程可参见文献[7]，因此本文研究的重点是使用粗糙集对抽取的记录进行属性约简。软件开发过程的不同阶段存在着不同种类的风险，按照项目管理的结果一般可以分为：进度风险，成本风险和质量风险，当然在不同的开发阶段各类风险关注的影响因素也不尽相同，本文抽取软件开发过程中影响开发进度的 8 条记录(X1,X2…X8)，每条记录有 5 个影响因素(团队凝聚力(TM),CMM级别©,技术成熟度(T),需求分析的质量(R)，个人能力(P))，开发进度(D)作为结果，将影响因素定义为 3 个级别：低(1)，中(2)，高(3)，结果定义为 3 个级别：延迟(1)，适中(2)和超前(3)，其中，值为 0~0.35 定义为 1，值为 0.36~0.7 定义为 2，值为 0.71~1 定义为 3，对属性进行离散化后结果如表 1 所示：

表 1　离散后的数据

U	TM	C	T	R	P	D
X1	1	1	1	1	1	1
X2	1	1	1	1	2	1
X3	2	2	2	2	1	2
X4	2	2	2	2	2	2
X5	2	3	2	2	1	2
X6	2	3	3	3	2	3
X7	1	1	1	1	2	1
X8	1	1	1	1	3	1

U/TM ={{x1,x2,x7,x8},{x3,x4},{x5,x6}}

U/C={{x1,x2,x7,x8},{x3,x4},{x5,x6}}

U/T={{x1,x2,x7,x8},{x3,x4 ,x5},{x6}}

U/R={{x1,x2,x7,x8},{x3,x4},{x5,x6}}

U/P={{x1,x3,x5},{x2,x4,x6,x7},{x8}}

U/D={{x1,x2,x7,x8},{x3,x4 ,x5},{x6}}

压缩后的简表为：

U/D	TM	C	T	R	P
{x1,x2,x7,x8}	1	1	1	1	2
{x3,x4,x5}	2	2	2	2	1
{x6}	2	3	3	3	2

可以验证信息表中 P 属性是多余的，可得到四个最简属性{TM,C,T,R}。然后根据因素间的影响关系建立贝叶斯网络模型如下：

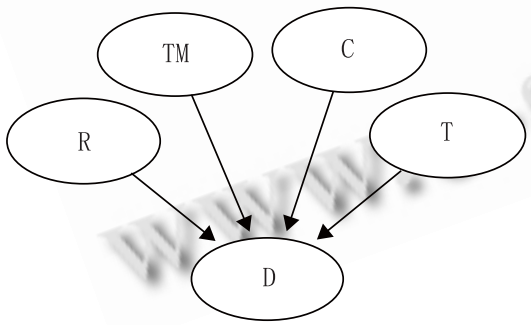

图 1　贝叶斯风险评估模型

根据建立的贝叶斯网络对影响开发进度的风险进行定量的评估，作为开发过程中制定开发进度计划的依据。在笔者开发的房产管理信息系统中，采用贝叶斯网络对需求进度评估得出需求分析需要 23 天，改进的贝叶斯网络进行评估得出需要 21 天，实际开发过程需求分析进行了 20 天，因此，改进的贝叶斯网络与实际情况更符合。

4　总结

在粗糙集对属性约简的基础上，贝叶斯网络以其强大的图形表示方法及推理理论，对软件开发中存在的风险因果关系进行建模，降低了无关因素对评估结果的影响，为制定合理的风险管理计划提供了依据。若能在进行贝叶斯网络建模及其推理过程中考虑到时间因素的影响，那么，风险评估的效果会更好，对于与时间相关的动态贝叶斯网络，还有待进一步研究。

参考文献

1 Williams RC, Pandelios GJ, Behrens SG. Software Risk Evaluation(SRE) Method Description(Version 2.0).1999.

2 Rangarajan A, Coughlan J, Yuille AL. A Bayesian Network Frame for Relational Shape Matching.IEEE Computer Society, 2003.

3 Jensen FV. Bayesian Networks and Decision Graphs. Springer-Verlag, New York, 2001.

4 Chickering DM. Learning equivalence classes of Bayesian Network Structs. Journal of Machine Learning Research, 2002

5 张文修,等.基于粗糙集的不确定决策.北京:清华大学出版社, 2005.

6 张文修,等.粗糙集理论与方法.北京:科学出版社, 2001.

7 姚全珠,任雪利,丁晓剑.基于贝叶斯网络的计划评审技术.西安理工大学学报, 2006,(22):137－140.