基于粗糙集与证据理论的测试用例优化研究*

孙青青，李龙澍，李学俊，徐 怡
（安徽大学 计算机科学与技术学院，合肥 230039）

摘 要：由于软件体系复杂度和数量不断增加使得测试用例的设计和选择越来越困难，为了能够在来自不同信息源的众多测试用例中选择有效的用例集，提出了一种基于粗糙集与证据理论的测试用例优化方法。粗糙集是一种处理不确定信息的有效方法，其本身不需要先验知识、有经典的约简算法等；D-S 理论处理来自不同信息源的数据合成问题，是经典概率论的一种扩展，但其本身存在许多不足。粗糙集的特点恰好可以弥补其不足，两者结合从而达到良好的效果。在实际应用分析中，证明该方法是可行而且有效的。

关键词：测试用例；粗糙集；证据理论；多源信息；约简
中图分类号：TP311 **文献标志码**：A **文章编号**：1001-3695(2012)07-2534-03
doi:10.3969/j.issn.1001-3695.2012.07.035

Research on test case optimization based on rough set and D-S theory

SUN Qing-qing, LI Long-shu, LI Xue-jun, XU Yi
(*School of Computer Science & Technology, Anhui University, Hefei* 230039, *China*)

Abstract: Because of software system complexity and a growing number of test cases, the designing and choosing of test cases are more and more diffcult and this paper put forward a kind of test cases optimization method for test cases synthesis optimization to select effective test cases from different sources. Rough set is a kind of effective method dealing with uncertain information, it doesn't need advantages of prior knowledge. D-S theory can deal with the data synthesis problem from different sources, and it is a classic extension of the theory of probability, but it has many shortcomings. However the characteristics of rough set just can make up the deficiencies, combine the two methods to achieve good effect. In the actual application analysis, this method is feasible and effective in the actual application analysis.

Key words: test case; rough set; D-S theory; multi-source information; reduction

测试用例的设计与优化是软件测试关键步骤之一。自动化测试是当前软件测试研究的热点之一，旨在为了节省人力、时间或者硬件资源，故如何优化和选择来自不同信息源的测试脚本是自动化测试的一个重要工作。

粗糙集作为一种新的处理不确定问题的理论[1]，是数据挖掘的有效方法之一。它的重要特点是不需要任何知识，只利用数据本身提供的信息，在保持知识的分类能力不变的前提下，通过知识约简，导出分类规则，提供决策依据。

证据(D-S)理论是用来处理由认识的局限性所导致的不确定性问题的数学工具[2]，处理来自不同信息源的数据合成问题，是经典概率论的一种扩展。但其存在很多缺点，主要表现在焦元爆炸、基本概率的指派由专家指派的主观性等。由于粗糙集本身的优势[3-5]，与 D-S 理论方法的结合可以有效克服上述问题。本文主要结合两者对应关系提出了一种基于测试用例的算法。

1 粗糙集理论

近年来，粗糙集在数据挖掘、人工智能、模式识别等领域得到了广泛的应用。

定义 1 设 $K=(U,R)$ 是一个近似空间，且依赖于知识 P，记做 $P\Rightarrow Q$，当且仅当：

$$k=\gamma_p(Q)=\frac{\mathrm{card}(\mathrm{POS}_p(Q))}{\mathrm{card}(U)}$$

以上定义了属性集的依赖度，有别于属性依赖度[6]。属性集依赖度的求解方法为 $k=\{$元素数$|\{$条件属性集等价类$\cap$决策属性集等价类$\}\}/\{$对象个数$\}$。

基于属性依赖度的决策分解算法[7]：

输入：决策信息系统 S 以及系统数量 N；

输出：分解后的子系统。

a)计算各个属性对决策属性的依赖度，构造相关矩阵；

b)将相关矩阵转换为对称矩阵，进一步转换为三角矩阵 D_{ij}；

c)while 条件属性的总标类数大于 N

从 D_{ij} 中选择最大的 d_{ij}；

if c_i 或者 c_j 已经归为某新类，THEN 另一个属性也归为该类，删除该属性的标号，并将 D_{ij} 的值置为 -1；

else if c_i 和 c_j 分别归于不同的类，则合并这两个类，删除其中一个的标号，并将 D_{ij} 的值置为 -1；

else c_i 和 c_j 归于一个类，并将 D_{ij} 的值置为 -1；

end if

收稿日期：2011-12-15；**修回日期**：2012-01-26 **基金项目**：安徽省自然科学基金资助项目(090412054)；安徽省重大科技专项资助项目(08010201002)；安徽高等学校省级自然科学基金资助项目(KJ2011Z020)

作者简介：孙青青(1987-)，女，硕士研究生，主要研究方向为粗糙集理论、软件测试(sqq080@126.com)；李龙澍(1956-)，男，教授，博导，主要研究方向为知识工程、软件分析与测试；李学俊(1976-)，男，副教授，主要研究方向为数据挖掘、数据库；徐怡(1981-)，女，副教授，主要研究方向为不精确信息处理.

end if

end while

d)根据聚类结果分解原始的决策表,得到属性分解后的各个子决策表;

e)算法结束。

约简主要分为属性约简与值约简,本文算法中使用的是快速约简算法[7],不用预先求出核,直接根据属性组合计算依赖度,使得当前属性集的依赖度与原来的决策表的相等,既保证了分类能力不变的情况下,有效地缩小了搜索空间,又提高了约简效率。算法如下:

输入:决策表 $S=\langle U,A,V,f\rangle$,其中 $A=C\cup D$;

输出:决策信息系统的最小约简。

a)设初始属性集 R 为空;

b)$T=R$;

c)$\forall x\in(C-R)$;

d)if $\gamma_{R\cup\{x\}}>\gamma_T(D)$,那么 $T=R\cup\{x\}$;

e)$T=R$;

f)until $\gamma_R(D)=\gamma_C(D)$,否则返回 c);

g)return R(最终返回的属性约简集合);

h)结束。

2　D-S 理论

Dempster 合成规则可描述为以下定理[2]:

定理 1　设 Bel_1,Bel_2 是同一识别框架 Θ 上的两个信任函数,基本可信度分配为 m_1 与 m_2,焦元分别为 $A_1,\cdots,A_K$ 和 $B_1,\cdots,B_L$,如果 $\sum_{A_i\cap B_j=A}m_1(A_i)m_2(B_j)<1$ 那么,函数 $m:2^\Theta\to[0,1]$ 对于所有的非空集合 $A\subseteq\Theta$ 满足 $m(\varnothing)=0$ 且

$$m(A)=\frac{\sum_{A_i\cap B_j=A}m_1(A_i)m_2(B_j)}{1-\sum_{A_i\cap B_j=\varnothing}m_1(A_i)m_2(B_j)}=$$

$$\frac{1}{K}\sum_{A_i\cap B_j=A}m_1(A_i)m_2(B_j),A=\varnothing$$

其中:$K=1-\sum_{A_i\cap B_j=\varnothing}m_1(A_i)m_2(B_j)>0$。

3　粗糙集与 D-S 理论

粗糙集与证据理论都是处理不确定信息的有力工具,且两者之间存在紧密的关系,表现在证据理论中的基本概率指派、信任函数、似然函数都可以用粗糙集的方法计算[8]。

定义 2　R 是全集 U 上的一个等价关系,其相应地划分为 $U/R=\{Y_1,Y_2,\cdots,Y_n\}$,$X$ 为 U 上的一个子集即 $X\subseteq U$,定义由 R 导出的 X 的下近似的质量为一个 $2^U\to[0,1]$ 的函数 Bel 如下:

$$\mathrm{Bel}(X)=\frac{|U_{Y_i\in U/R,Y_i\subseteq X}Y_i|}{|U|}=\frac{|R_-(X)|}{|U|}$$

定义 3　R 是全集 U 上的一个等价关系,其相应地划分为 $U/R=\{Y_1,Y_2,\cdots,Y_n\}$,$X$ 为 U 上的一个子集即 $X\subseteq U$,定义由 R 导出的 X 的上近似的质量为一个 $2^U\to[0,1]$ 的函数 Pls 如下:

$$\mathrm{Pls}(X)=\frac{|U_{Y_i\in U/R,Y_i\neq\varnothing}Y_i|}{|U|}=\frac{|R^-(X)|}{|U|}$$

由以上定义可以进一步建立粗糙集和证据理论的联系,证据合成中的证据来源于各个子决策系统即多源问题。粗糙集中的近似空间、边界域、上下近似分别对应证据理论中的辨识框架、信任函数、似然函数、置信度等概念,由此建立两者之间的联系。以下定理准确说明了两者之间的关系。

定义 4　决策信息系统 $S=\langle U,A=C\cup D,V=V_C\cup V_D\rangle$,其中 C 为条件属性,D 为决策属性。决策划分为 $U/D=\{X_1,X_2,\cdots,X_n\}$,其中 $X_i=\{x\in U|\alpha(x)=d_i\in V_D\}$,$\alpha$ 是映射函数,决策属性 D 的取值空间为 $V_D=\{d_1,\cdots,d_n\}$,$\beta\subseteq V_D$,则有:

a)$m_C(\varnothing)=0$;

b)$m_A(\beta)=|\{x\in[x]_{\mathrm{IND}(C)}:\alpha_C([x]_{\mathrm{IND}(C)})=\beta\}|/|U|$;

c)$\mathrm{Bel}(\beta)=\mathrm{apr}_\beta=|C_-(U_{i\in\beta}X_i)|/|U|$,其中 apr_β 是广义 β 下近似质量函数;

(4)$\mathrm{Pl}(\beta)=\mathrm{apr}^\beta=|C^-(U_{i\in\beta}X_i)|/|U|$,其中 apr^β 是广义 β 上近似质量函数。

至此已经明确得到了粗糙集理论与证据理论的相互联系,从而为下面基于粗糙集与证据理论对于多源测试用例的优化奠定了基础。

4　基于粗糙集与 D-S 的测试用例优化

自动化测试是近年来软件测试研究热点之一,众多国内外专家与学者可谓是硕果累累,其旨在节省人力、时间或者硬件资源。众所周知,测试脚本优化与选择在软件测试中占据着举足轻重的地位,其目标在于以最少的测试用例达到相同的覆盖率及测试结果,如性能、配置等测试需要大数据量输入的情况下而用自动化测试来替代人工测试[9]。

自动化测试脚本的一个重要特点是它的可重复性,如在回归测试中重复单一的数据录入或是击键等测试操作造成了不必要的时间浪费和人力浪费[9]。由于测试脚本的重复使用,存在以下两个问题:

a)不同测试需求,要求测试脚本关注的参数不同,如何依据客观现实来选择所需的最优测试脚本;例如在回归测试中软件测试人员关注的是变更函数数量、覆盖率等,而在压力测试中测试人员关注的是人数等因素。

b)测试脚本重复使用,不同的使用环境及人员得到的相关性能参数是不同的,如何利用已有记录而得到的评价参数服务于下一次的测试中。

针对以上两个问题,本文提出了以下的解决方案:

粗糙集理论中属性集依赖度的概念,在聚类的过程中以往的观念是将每个属性平等对待,而在测试过程中更重要的是某两个或是更多的参数对测试用例的参考价值。故而可以利用粗糙集属性集依赖度的思想将所关注的属性聚类过程中看做一个整体求解依赖度,即定义 1 中 k,实践证明该方法有效地降低了聚类的时间复杂度。

D-S 理论的合成规则可以较好地解决问题 b)。各个时间段、人员、环境使用下的测试脚本结果作为不同的信息源。由于证据理论在信息融合中本身的缺陷,利用粗糙集理论预先对数据进行聚类以及约简处理,不仅能够较好地弥补证据理论本身焦元爆炸及先验知识主观性的问题,而且能够降低海量数据本身的复杂度。

由以上论述可以得出测试用例优化算法如下:

输入:决策表 $S=\langle U,A,V,f\rangle$;

输出:证据的基本可信度。

a)从不同的阶段的测试用例使用结果收集大量的原始样本数据,对样本信息离散化处理,使之构成信息系统 $S=\langle U,A,V,f\rangle$;

b)依据定义 1 求解条件属性中各属性及属性集依赖度,

进行聚类;

c)对分解得到的各个子决策表,依据上文中介绍的快速约简算法进行属性约简;

d)优化后的各个子表,按照证据理论与粗糙集的关系即定理 2,求得各个子表的基本可信度分配值;

e)依据步骤 e)得到的可信度分配值,利用证据理论合成规则即定理 1 进行合成,得到合成优化后的可信度分配;

f)利用证据理论决策推理;

g)算法结束。

5 实例分析及结论

5.1 实例分析

基于以上算法,下面以一实例说明。决策信息系统 $S = \langle U, C \cup D, V = V_C \cup V_D \rangle$,如表 1 所示。

表 1 原始决策表

U	C_1	C_2	C_3	C_4	C_5	C_6	D
X_1	0	1	0	0	1	1	2
X_2	0	0	1	0	1	1	2
X_3	1	1	0	1	1	1	1
X_4	0	2	0	1	1	1	3
X_5	1	0	0	0	0	2	1
X_6	1	0	0	0	0	2	2
X_7	1	0	0	0	0	3	1
X_8	1	0	0	1	0	2	2
X_9	1	0	1	0	0	2	2
X_{10}	1	0	1	0	0	2	3
X_{11}	1	0	1	0	0	3	1
X_{12}	1	0	1	0	1	2	2
X_{13}	1	1	0	0	1	1	3

根据属性依赖度的思想聚类,聚类的个数在此设为 2,利用分解算法将属性集分解为决策表 S_1 的属性集 $C_1 = \{c_1, c_2, c_3, c_5, c_6\}$ 以及决策表 S_2 属性集 $C_2 = \{c_2, c_3, c_4, c_5, c_6\}$。分别对两个子决策表依据快速约简算法进行属性约简得到以下两个约简后的子决策表如表 2、3 所示。

表 2 约简后的 S_1 决策表

U	C_1	C_3	C_5	C_6	D
X_1	0	0	1	1	2
X_2	0	1	1	1	2
X_3	1	0	1	1	1
X_4	0	0	1	1	3
X_5	1	0	0	2	1
X_6	1	0	0	2	2
X_7	1	0	0	3	1
X_8	1	0	0	2	2
X_9	1	1	0	2	2
X_{10}	1	1	0	2	3
X_{11}	1	1	0	3	1
X_{12}	1	1	1	2	2

由定义 4 可知,广义决策值的划分就对应证据中的辨识框架,从而可计算得到各个证据元的基本概率分配。由表 2 得到的基本概率分配、信度函数值和似真度函数值如下:$m_{c_1}(1) = 1/12, m_{c_1}(2) = 1/12, m_{c_1}(3) = 0, m_{c_1}(1,2) = 1/6, m_{c_1}(2,3) = 1/6, m_{c_1}(1,3) = 1/6, m_{c_1}(1,2,3) = 1/4$。

同理可得由表 3 得到的基本概率分配、信度函数值和似真度函数值如下:$m_{c_2}(1) = 1/12, m_{c_2}(2) = 1/6, m_{c_2}(3) = 1/12, m_{c_2}(1,2) = 1/6, m_{c_2}(2,3) = 0, m_{c_2}(1,3) = 0, m_{c_2}(1,2,3) = 1/12$。

表 3 约简后的 S_2 决策表

U	C_2	C_3	C_5	C_6	D
X_1	1	0	1	1	2
X_2	0	1	1	1	2
X_3	1	0	1	1	1
X_4	2	0	1	1	3
X_5	0	0	0	2	1
X_6	0	0	0	2	2
X_7	0	0	0	3	1
X_8	0	0	0	2	2
X_9	0	1	0	2	2
X_{10}	0	1	0	2	3
X_{11}	0	1	0	3	1
X_{12}	0	1	1	2	2

由以上两组证据的可信度分配,根据证据合成规则得到 $m(1) = 20/129, m(2) = 38/129, m(3) = 7/129, m(3) = 7/129, m(2,3) = 12/129, m(1,3) = 12/129$。该实例表明,对 {1},{2},{3} 的支持度远高于其他几个结论。由此可知,粗糙集不需要先验知识避免了专家指定可信度分配带来的主观性,属性约简对海量数据进行约简而不改变原有数据的分类能力;粗糙集与证据理论的结合,扬长避短得到合成后可信度分配。由此根据合成后的可信服分配进行决策,即选择合适的测试用例。

5.2 结论

测试用例的优化是软件测试至关重要的部分,其质量直接决定了软件测试的质量。面对如今海量的数据信息,如何在保持覆盖率不变的情况下,减少测试用例数量以提高测试效率显得越来越重要。

本文运用粗糙集和证据理论,结合自动化测试中测试脚本的特点,提出了一种测试用例优化分析的方法。首先对历史数据进行聚类分解,得到不同的子决策系统,对各个子系统进行属性约简;根据粗糙集与 D-S 的对应关系求得各个子决策系统的基本信度分配,再由证据理论的合成规则求得合成后的可信度分配,由此进行决策,即选择合适的测试用例。

参考文献:

[1] LEI Hong-yan, TIAN Wang-lan, ZOU Han-bin. The study on data mining methods based on rough set theory and CART for incomplete data [C]//Proc of the 3rd Pacific-Asia Conference on Circuits, Communications and System. 2011:1-4.

[2] BASIR O, YUAN Xiao-hong. Engine fault diagnosis based on multi-senior information fusion using Dempster-Shafer evidence theory[J]. Information Fusion, 2007, 8(4):379-386.

[3] 张国军. 基于粗糙集的相对属性约简算法及决策方法研究[D]. 武汉:华中科技大学,2010.

[4] 王国胤,姚一豫,于洪. 粗糙集理论与应用研究综述[J]. 计算机学报,2009,32(7):1229-1246.

[5] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社,2001:26-32.

[6] 刘柳明,王加阳,罗安. 决策表属性分解的等价性研究[J]. 计算机应用研究,2007,24(8):67-69.

[7] 周勇. 基于粗糙集于证据理论的信息融合研究[D]. 长沙:中南大学,2008.

[8] 刘海燕,赵宗贵,刘熹. D-S 证据理论中冲突证据的合成方法[J]. 电子科技大学学报,2008,37(5):701-704.

[9] LEVESQUE M, LOUIE J, GUERREROA M. Test execution control tool:automating testing in spac-ecraft integration and test environments [C]//Proc of IEEE Aerospace Conference. USA:IEEE, 2000:389-395.