Software Engineering and Applications 软件工程与应用, 2013, 2, 47-54
http://dx.doi.org/10.12677/sea.2013.22009 Published Online April 2013 (http://www.hanspub.org/journal/sea.html)

Based on Rough Set Machine Learning of WLC Estimation Method*

Chenguang Jing[1], Xiaochen Duan[2]

[1]China Railway Siyuan Survey and Design Group Co., LTD., Wuhan
[2]Shijiazhuang Tiedao University, Shijiazhuang
Email: 13044964@sohu.com

Received: Mar. 11th, 2013; revised: Mar. 17th, 2013; accepted: Apr. 4th, 2013

Abstract: In this paper, rough set theory in knowledge discovery on the superiority of the combination of machine learning theory to the actual sample quantities, for example, the uncertainty of the historical data under the influence of life cycle cost estimation problem. In the light of the specific examples based on rough sets is given from the modeling, the effective data screening to decision rules generation, life cycle cost of the final results obtained the complete estimation process. This paper attempts to estimate life cycle cost of the introduction of rough set theory of machine learning, data from a large number of experimental works of the most influential factors in selection, the decision attribute and condition of maintaining the dependencies between attributes does not change the premise, according to engineering knowledge base to find the equivalence relations between the redundancy to simplify the decision table, to ensure that their classification ability, reduction factor out the weak links, and finally to study rough set decision rules are implemented cost forecast. Confusion matrix by cross-validation showed that the application of rough set theory under the influence of data uncertainty to resolve the full life cycle cost estimate is feasible.

Keywords: Whole Life Costing; Rough Set; Machine Learning

基于粗糙集机器学习的全生命周期造价估算方法研究*

景晨光[1]，段晓晨[2]

[1]中铁第四勘察设计院，武汉
[2]石家庄铁道大学，石家庄
Email: 13044964@sohu.com

收稿日期：2013 年 3 月 11 日；修回日期：2013 年 3 月 17 日；录用日期：2013 年 4 月 4 日

摘　要：本文利用粗糙集理论在知识发现上的优越性，结合机器学习的原理，以实际工程量清单样本为例，研究了历史数据不确定性影响下全生命周期造价的估算问题。在结合具体实例的基础上，给出了粗糙集从建模、有效数据筛选到决策规则生成、最终得出全生命周期造价结果的完整估算过程。本文尝试在全生命周期造价估算中引人粗糙集机器学习理论，从大量实测工程数据中优选出最有影响的因素，在保持决策属性和条件属性之间的依赖关系不变化的前提下，根据其等价关系寻找工程知识库中的冗余关系，从而简化决策表，确保其分类能力，约简掉联系较弱的因素，最后以粗糙集决策规则学习的形式实现造价预测。通过混淆矩阵交叉验证表明，应用粗糙集理论解决数据不确定性影响下的全生命周期造价估算是可行的。

关键词：全生命周期造价；粗糙集；机器学习

*基金项目：本课题的研究来源于国家自然科学基金项目。政府投资项目全面投资控制理论与方法研究(70373032)。

1. 引言

长期以来，我国一直把建设项目投资控制的重点放在施工阶段，认为施工阶段是建筑产品形成实体的最后阶段，这样做，尽管是必要的且取得一定的效果，但是对整个造价控制工作来讲，仍然具有片面性，存在明显的缺陷。对那些因缺乏科学的可行性研究，或投资估算不实，或设计不科学、不合理、脱离实际等所造成的拖延工期、投资浪费等事项只能发现，而不能及时纠正并消除偏差，更不能预防偏差的发生，只是被动的控制工程造价，是事后控制[1]。因此，从全过程造价到全生命周期造价管理思想的转变是目前我国工程造价管理领域急需解决的重要问题。由于，工程造价系统经历从模糊性(不确定)、渐进到确定的过程，前期不确定性的投资起着至关重要的控制作用。现行工程造价投资估算、决策、控制理论方法的线性、确定性、简单性、滞后性的缺陷，导致投资目标确定误差大(预测不准)和控制可靠性极不稳定(三超问题)，引发了现实中一系列的质量、工期，超支等问题，本文试图利用粗糙集机器学习算法进行全生命周期造价(WLC)估算，以期提高 WLC 估算水平。

2. 模型简介

2.1. 全生命周期造价(WLC)

全生命周期造价(whole life costing，WLC)指工程项目在整个生命周期内发生的费用[2]。是一种实现工程项目的全生命周期，包括可行性研究期、设计期、建设期、使用期、翻新与拆除期、弃置回收期等阶段总造价最小化的方法。指从该项目的设想、产生的过程及使用阶段到该产品的废弃以及弃置回收这一整个的生命周期。从工程项目的全生命周期角度出发去考虑问题，既考虑建设期的各个阶段，同时也考虑了运营维护和弃置回收等阶段的研究，按照全生命周期成本最低的思想进行工程设计和投资优化。它考虑的时间范围更长，也更合理。综合考虑项目的建设、运营维护、弃置回收等成本，按照生命周期成本最小化的原则，指定最佳投资方案，从而更加科学合理的实现投资估算。

2.2. 粗糙集(RS)模型

粗糙集(Rough set, RS)理论作为人工智能领域中的一种新方法[3]，其优点是无需提供问题所需处理的数据集合之外的任何先验知识，其模拟人类抽象逻辑思维[4]，完全从数据中得到规律或其它结论，真正实现了“让数据自己说话”。目前，粗糙集理论与神经网络、演化计算、模糊系统及混沌系统一起被公认为人工智能的五大新兴技术[5]。其模型如下：

设集合 $X \subseteq U$， R 是一个等价关系，称 $\underline{R}X = \{x \mid x \in U, \text{且} [x]_R \subseteq X\}$ 为集合 X 的 R 下近似集；称 $\overline{R}X = \{x \mid x \in U, \text{且} [x]_R \cap X \neq \Phi\}$ 为集合 X 的 R 上近似集。称集合 $\mathbf{BN}_R(X) = \overline{R}X - \underline{R}X$ 为 X 的 R 边界域；称 $\mathbf{POS}_R(X) = \underline{R}X$ 为 X 的 R 正域；称 $\mathbf{NEG}_R(X) = U - \overline{R}X$ 为 X 的 R 负域。用图 1 表示更加直观：

当 $\mathbf{BN}_R(X) = \Phi$ 时，即 $\overline{R}X = \underline{R}X$，称 X 是 R 精确集；当 $\mathbf{BN}_R(X) \neq \Phi$ 时，即 $\overline{R}X \neq \underline{R}X$，称 X 是 R 粗糙集。集合 X 的近似精度为：$\alpha_R(X) = \dfrac{|\underline{R}X|}{|\overline{R}X|}$，$(X \neq \Phi)$，$|X|$ 为集合 X 的基数，当 $0 \le \alpha_R(X) < 1$ 时，X 是 R 的粗糙集。

3. 粗糙集机器学习估算建模

基于粗糙集的学习模型详见图 2，其过程如下：

设决策表：

$S = (U, A, V_A, f), C \cup D = A, C \cap D = \varnothing$，

$C = \{c_1, c_2, \cdots, c_k\}$ 为条件属性集，D 为决策属性集。

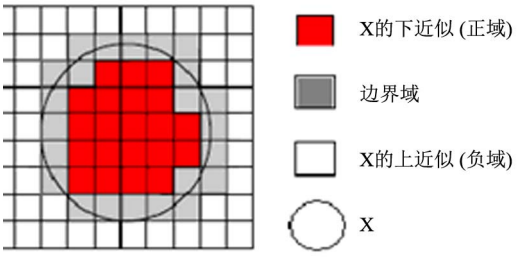


Figure 1. RS schematic diagram
图 1. RS 示意图

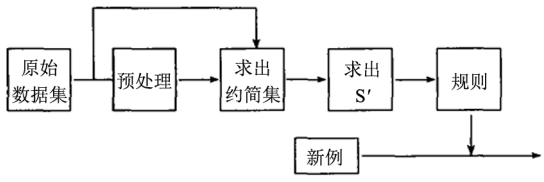


Figure 2. Rough set model of machine learning
图 2. 粗糙集机器学习模型

1) 首先对原始决策表进行数据预处理：补齐，离散化，完备化，相容性检验，约简；否则 go to 2)；

2) 利用粗糙集约简算法，化简该表，求约简集，得出一个新的决策表，记为S'，并求出规则集；

3) 如果有新例加入本系统，先进行预处理，然后利用 2)中求出的决策规则，进行学习；

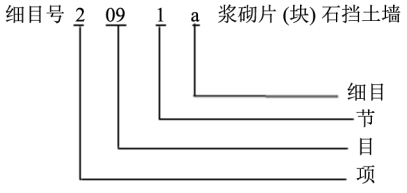
4) 如果学习质量没有满足用户或系统要求(如给出阈值)，go to 1)重新处理，直到满足要求为止；

5) 利用混淆矩阵(Confusion matrix)进行验证。利用上述的粗糙集机器学习模型，本文将原始决策表分割(Split)成 2 个子表，一个作为训练子表，一个作为学习子表，即图 2 中的”粗糙集机器学习模型”，作为学习样本，分割方法如下：

原信息系统(IS)为$S=(U,A)$，分为$S_1=(U_1,A)$，$S_2=(U_2,A)$，两个子表，且满足如下条件：$U_1\cup U_2=U$，$U_1\cap U_2=\varnothing$。

4. 实例与仿真

本文以高速公路的路基工程为例，依据交通部《公路工程国内招标文件范本》[6]和建设部《建设工程工程量清单计价规范》编制而成的公路工程工程量清单计量规则为依据，将工程量清单中的项目以相应规范中的条目进行标注，力求做到标准化、统一化。通过对历史路基工程的工程量清单数据的收集、整理，分析出不同工程特征数据。公路工程工程量清单计量规则项目号的编写分别按项、目、节、细目表达，工程量清单细目号对应方式示例如下：



为了更好的结合粗糙集(RS)进行分析，考虑到粗糙集自身的特点，如选用“目”作为工程特征类目，则范围太大，即便是通过粗糙集提取出了工程特征，也失去了对之后作为类似工程选择依据的意义，如选用“细目”作为工程特征类目，则范围太狭隘，很可能以此为依据而找不到任何类似工程，在充分考虑了工程造价工程量清单计价模式和粗糙集(RS)的特点后选择以工程量清单细目中的“节”为工程特征类目。

工程特征属性提取的意义在于准确客观的计算影响造价的各个“节”特征在系统内的影响程度。克服传统工程特征专家经验确定方法的主观性，使得工程特征类目的选取更具客观性，从粗糙集的角度去分析即去除多余的属性或者干扰属性。如果对属性不做选择，会带来许多问题。首先，在全生命周期造价的实施过程中，实际的数据挖掘和信息融合系统所处理的数据是非常庞杂的，从海量数据中挖掘出有用的信息是比较困难的；并且，大量的数据所携带的信息并不都是对决策有用的。因此，去掉无关的和冗余的数据，筛选出有用的数据是数据处理很重要的步骤[6]；而且，多余的属性或者干扰属性会增加分类算法的复杂性，有时会导致以工程特征为依据寻找类似工程的效果急剧下降[7,8]。

4.1. 原始连续量决策信息表的获取和预处理

因为工程量清单属于保密资料，所以本文仅将工程量清单进行编号，而不列出实际公路施工段名称。按造价指数和地区调整系数对数据消除时间、地区等差别。运营维护成本将按初始成本所占总造价的比例，进行分摊与合并，运营维护成本的计入反映了全生命周期造价的理念，经调整计算后得到的工程实例数据信息表详见表 1。

4.2. 对原始的决策信息表 M 进行离散化处理

本节使用等频离散化方法进行离散计算，其方法如下：

设在值域V_a上有 n 个原始离散值，由给定参数 k 把这 n 个离散值分成 k+1 段，每段有 $n/(k+1)$个离散值，则断点c_j^a的确定方法如下：对于任意属性 a，以值升序集合表示它的值域，即

$$\begin{cases} a(U)=\{v_1^a,\cdots,v_m^a,v_{m+1}^a,\cdots,v_n^a\}, \\ v_1^a=c_0^a,\ \ v_n^a=c_{k+1}^a,\ \ a\in A, i=1,2,\cdots,l. \end{cases}$$

设断点c_j^a落入值v_m^a和值v_{m+1}^a之间，则断点$c_j^a=\left(v_m^a+v_{m+1}^a\right)/2$，这样可求出值域$V_a$中的 k 个断点$\left(c_1^a,c_2^a,\cdots,c_k^a\right)$。于是$V_a$就被划分为 k + 1 个区间，即P_a，每个区间用一个离散值来代替。如果V_a为决策属性的值域，则同样可以用区间重心法或区间值平均法求出各决策区间的离散值。计算结果如表 2 所示。

Table 1. Engineering instance data
表 1. 工程实例数据

工程序号	清理与掘除	挖除旧路面	…	挂网喷混植生防护	二布一膜复合土工膜	挡土墙	预应力锚索	浆砌片石锥(护)坡	总造价
1	30,210	0	…	0	20,943	1,923,340	0	0	16,461,182
2	0	0	…	0	0	166,334	0	0	372,997
3	236,127	3152	…	280,644	0	1,220,887	0	179,321	16,851,358
4	219631.2	0	…	108,359	0	19361.2	0	289718.3	9,196,058
5	44,762	0	…	0	0	0	0	0	6,665,851
…	…	…	…	…	…	…	…	…	…
12	2654	0	…	0	0	11,409	0	0	789,856
13	546,251	282,302	…	0	0	0	0	0	19,815,306

Table 2. Quantity of discrete values
表 2. 工程量清单离散值

工程量清单目	属性离散码		
	1	2	3
清理与掘除	[*, 13309.5)	[13309.5, 308330.0)	[308330.0, *)
挖除旧路面	[*, 472.13)	[472.13, 2382.79)	[2382.79, *)
拆除结构物	[*, 450.50)	[450.50, 29649.60)	[29649.60, *)
路基挖方	[*, 635531.00)	[635531.00, 2213240.00)	[2213240.00, *)
改路、改河、改渠挖方	[*, 7807.9)	[7807.9, 73807.8)	[73807.8, *)
借土挖方	0	11603	
路基填筑	[*, 555958.00)	[555958.00, 2035040.00)	[2035040.00, *)
改路、改河、改渠填筑	[*, 592)	[592, 152617)	[152617, *)
结构物台背回填及锥坡填筑	[*, 64459.5)	[64459.5, 727665.0)	[727665.0, *)
软土地基处理	[*, 79734.20)	[79734.20, 1216550.00)	[1216550.00, *)
滑坡处理	0	443808	
填挖交界处治	0	139642.5	
黄土处理	[*, 19634)	[19634, 113904)	[113904, *)
过湿润土处理	[*, 97965.7)	[97965.7, 3852710.0)	[3852710.0, *)
边沟	[*, 10651.00)	[10651.00, 511644.00)	[511644.00, *)
排水沟	[*, 235805.0)	[235805.0, 911194.0)	[911194.0, *)
截水沟	[*, 210.00)	[210.00, 27574.20)	[27574.20, *)
砌片石急流槽	[*, 8387.69)	[8387.69, 45519.00)	[45519.00, *)
路基渗(盲)沟	[*, 5098.50)	[5098.50, 168486.00)	[168486.00, *)
涵洞上下游改沟、铺砌	[*, 105)	[105, 129234)	[129234, *)
通道排水	[*, 2196.5)	[2196.5, 16299.0)	[16299.0, *)
坡面植物防护	[*, 39213)	[39213, 132408)	[132408, *)
浆砌片石护坡	[*, 12642.50)	[12642.50, 880475.06)	[880475.06, *)
预制混凝土块护坡	[*, 256047.00)	[256047.00, 663596.00)	[663596.00, *)
护面墙	[*, 5367.5)	[5367.5, 284289.0)	[284289.0, *)
土工网	[*, 20999.20)	[20999.20, 345649.00)	[345649.00, *)
挂网喷混植生防护	[*, 54180)	[54180, 194502)	[194502, *)
二布一膜复合土工膜	[*, 10472)	[10472, 104152)	[104152, *)
挡土墙	[*, 5990.0)	[5990.0, 414083.0)	[414083.0, *)
预应力锚索	0	437150	
浆砌片石锥(护)坡	[*, 89660.5)	[89660.5, 234520.0)	[234520.0, *)
总造价	[*, 5462840)	[5462840, 18333300)	[18333300, *)

4.3. 决策表的确定

最终得到离散化属性表如图 3 所示。其中 a 代表清理与掘除，b 代表挖除旧路面，c 代表拆除结构物，d 代表路基挖方，…，ae 代表浆砌片石锥(护)坡，D 代表项目全生命周期造价。

4.4. 决策表的完备化处理

根据实际情况，由于决策表中的数据均来自实际的工程量清单，对每个对象下的所有属性及属性值经过整理、分析、处理，信息表中没有空值存在，都是已知的，该原始决策信息表是完备的，则可省略决策表的完备化处理这一步。

4.5. 判断决策表的相容性

本文应用 $poc_C(D)=U$ 来判断决策表的相容性。即如果决策属性对条件属性的下近似等于全集时，该决策表是相容的，否则是不相容的。

由图 3 离散化表得

$$\mathrm{ind}(C)=\{\{1\}\{2\}\{3\}\{4\}\{5\}\{6\}\{7\}\{8\}\{9\}\{10\}\{11\}\{12\}\{13\}\{14\}\}$$

$$\mathrm{ind}(D)=\{\{1,3,4,5,10\}\{2,6,8,11,12\}\{7,9,13,14\}\}$$

则

$$pos_C(D)=\underset{\forall x\in ind(D)}{U}\underline{C}(x)$$
$$=\{\{1\}\cup\{2\}\cup\{3\}\cup\{4\}\cup\{5\}\cup\{6\}\cup\{7\}$$
$$\cup\{8\}\cup\{9\}\cup\{10\}\cup\{11\}\cup\{12\}\cup\{13\}\cup\{14\}\}=U$$

即该表为相容决策表。

4.6. 将原始决策表分为两个子表

以实际工程量清单经过粗糙集预处理后得到的图 3 为例，我们首先把该表分成两个子表(Subtable)，子表(1)和子表(2)，详见图 4 和图 5。

U	a	b	c	d	e	⋮	G	h	i	⋮	z	aa	ab	ac	ad	ae	D
1	2	1	1	1	1	⋮	3	1	2	⋮	1	1	2	3	1	1	2
2	1	1	1	1	1	⋮	1	1	1	⋮	1	1	1	2	1	1	1
3	2	3	1	2	2	⋮	3	2	1	⋮	3	3	1	3	1	2	2
4	2	1	1	2	1	⋮	2	1	1	⋮	2	2	1	2	1	3	2
5	2	1	1	2	2	⋮	2	2	2	⋮	1	1	1	1	1	1	2
6	1	1	1	1	1	⋮	2	1	1	⋮	1	1	1	1	1	1	1
7	3	2	3	2	3	⋮	3	3	1	⋮	2	1	1	3	1	1	3
8	3	1	2	2	2	⋮	2	3	1	⋮	1	1	1	1	2	1	1
9	3	2	3	3	3	⋮	3	3	3	⋮	1	1	1	2	1	1	3
10	2	1	1	3	2	⋮	2	2	1	⋮	1	1	3	2	1	1	2
11	1	1	1	1	2	⋮	1	2	1	⋮	1	1	1	3	1	1	1
12	1	1	2	1	1	⋮	1	1	1	⋮	1	1	1	2	1	1	1
13	3	3	2	3	3	⋮	1	3	3	⋮	3	1	1	1	1	1	3
14	1	1	1	3	3	⋮	1	2	1	⋮	1	1	1	1	1	1	3

Figure 3. discretization table
图 3. 离散化表

U	a	b	c	d	e	f	g	⋮	s	t	u	v	w	x	y	z	aa	ab	ac	ad	ae	D
1	2	3	1	2	2	1	3	⋮	1	1	1	3	3	1	3	3	3	1	3	1	2	2
2	2	1	1	2	2	1	2	⋮	1	1	3	1	1	1	3	1	1	1	1	1	1	2
3	3	1	2	2	2	1	2	⋮	1	2	1	3	2	1	2	1	1	1	1	2	1	1
4	3	2	3	3	3	1	3	⋮	3	1	1	1	3	3	2	1	1	1	2	1	1	3
5	1	1	2	1	1	1	1	⋮	1	1	1	1	1	1	1	1	1	1	2	1	1	1
6	3	3	2	3	3	1	1	⋮	3	3	1	1	3	2	2	3	1	1	1	1	1	3
7	1	1	1	3	3	1	1	⋮	2	2	1	1	1	1	1	1	1	1	1	1	1	3

Figure 4. Child table (1)
图 4. 子表(1)

U	a	b	c	d	e	f	g	⋮	s	t	u	v	w	x	y	z	aa	ab	ac	ad	ae	D
1	2	1	1	1	1	1	3	⋮	2	1	2	1	3	1	1	1	1	2	3	1	1	2
2	1	1	1	1	1	1	1	⋮	1	1	1	1	1	1	1	1	1	1	2	1	1	1
3	2	1	1	2	1	1	2	⋮	1	1	2	2	2	1	3	2	2	1	2	1	3	2
4	1	1	1	1	1	1	2	⋮	1	1	1	2	2	1	2	1	1	1	1	1	1	1
5	3	2	3	2	3	1	3	⋮	3	1	1	1	2	2	1	2	1	1	3	1	1	3
6	2	1	1	3	2	2	2	⋮	2	1	3	1	2	1	1	1	1	3	2	1	1	2
7	1	1	1	1	2	1	1	⋮	2	1	1	1	1	1	1	1	1	1	3	1	1	1

Figure 5. Child table (2)
图 5. 子表(2)

4.7. 求离散的决策信息表 $dM_j\left(j=1,\cdots,q\right)$ 条件属性的约简提取规则集

根据 RS 属性约简算法，在条件属性集中可以去掉那些对决策属性没有影响或相对影响较小的属性，从而杜绝了由传统凭主观经验选择工程特征类目不科学的弊端。求出决策信息表中第 j 个条件属性相对于第 j 个决策属性 d_j 的相对约简，去掉那些对决策属性 d_j 而言不必要和次要的条件属性，于是就将第 j 个决策信息表约简成决策表，即

$$dM'_j=\left(U,C',d_j,dV_{C'},dV_{d_j},df'_j\right)$$

式中 $C'\subseteq C$， $j=1,\cdots,q$。由于实例中的决策属性是单一的，在此， $j=1$。

在粗糙集软件 Rosetta 中，选择 Genetic Algorithm，对属性进行约简，通过粗糙集约简，去除在本系统中冗余的属性，提取出有价值的，能代表本系统特征的数据，最后得到两个子表的约简和子表(1)的规则集如表 3，表 4 所示。

上述决策规则：如 F8(3) AND F22(1) AND F25(2) => F32(3)，根据图 2 实例工程的数据信息表和离散图 3 对照，可以用如下语言来解释：当改路、改河、改渠填筑造价为 652,795，坡面植物防护造价为 0，护面墙造价为 86,236，则全生命周期造价为 21294604.27 元，其他规则依次类推。子表(1)的约简集部分截图和子表(2)的规则集部分截图详见图 6 和图 7。

4.8. 造价估算与检验

在计算全生命周期造价时，为了进一步验证算法所得约简集对分类的影响，下面采用交叉验证(cross validation)进行算法验证。交叉验证是机器学习中广泛使用的一种技术，它可以估计出一种分类方法的预测准确率[7]。交叉验证法将数据中的一部分作为训练数据训练出分类器，将训练出的分类器在其余数据上作测试，得出的准确率作为对实际准确率的估计。论文将收集到的路基工程量清单总表分为两部分，第一部分(子表 1)用来进行属性的约简和规则集的计算，第二部分(子表 2)用来进行模型的学习与估算检验。图 8 给出了实验系统中机器学习新例情况的粗糙集混淆矩阵。

Table 3. Two child table
表 3. 两个子表的约简

Number	子表(1) Reduct	Number	子表(2) Reduct
1	{F3, F20}	1	{F1}
…	…	…	…
60	{F8, F10, F16}	35	{F4, F28, F31}
61	{F3, F14, F16}	36	{F7, F14, F28}
…	…	…	…
99	{F3, F9, F27}	48	{F16, F21, F28}
…	…	…	…
123	{F8, F10, F19}	57	{F2, F6, F21}

Table 4. Child table (1) set of rules
表 4. 子表(1)的规则集

Number	Rule	RHS Support	RHS Accuracy	LHS Coverage	RHS Coverage
1	F3(1) AND F20(1) => F32(2)	2	1.0	0.285714	1.0
2	F8(2) AND F20(1) => F32(2)	2	1.0	0.285714	1.0
…	…	…	…	…	…
25	F4(2) AND F30(1) => F32(2)	2	1.0	0.285714	1.0
26	F4(3) AND F30(1) => F32(3)	2	1.0	0.428571	1.0
…	…	…	…	…	…
50	F11(1) AND F16(1) AND F27(1) => F32(1)	2	1.0	0.285714	1.0
51	F11(1) AND F16(3) AND F27(1) => F32(3)	2	1.0	0.285714	0.666667
…	…	…	…	…	…
74	F8(3) AND F22(1) AND F25(2) => F32(3)	2	1.0	0.285714	0.666667

	Reduct	Support	Length
1	{F3, F20}	100	2
2	{F8, F20}	100	2
3	{F3, F10}	100	2
4	{F5, F10}	100	2
5	{F3, F19}	100	2
6	{F7, F10}	100	2
7	{F3, F5}	100	2
8	{F4, F10}	100	2
9	{F20, F25}	100	2
10	{F7, F18}	100	2
11	{F18, F19}	100	2
12	{F8, F17}	100	2
13	{F16, F29}	100	2
14	{F4, F23}	100	2
15	{F7, F8}	100	2
16	{F7, F15}	100	2
17	{F4, F8}	100	2
18	{F10, F15}	100	2
19	{F2, F15}	100	2
20	{F4, F25}	100	2

Figure 6. Child table (1) reduction set
图 6. 子表(1)约简集

	Rule
1	F1(2) => F32(2)
2	F1(1) => F32(1)
3	F1(3) => F32(3)
4	F17(2) AND F21(2) => F32(2)
5	F17(1) AND F21(1) => F32(1)
6	F17(3) AND F21(1) => F32(3)
7	F17(3) AND F21(3) => F32(2)
8	F23(3) AND F29(3) => F32(2)
9	F23(1) AND F29(2) => F32(1)
10	F23(2) AND F29(2) => F32(2)
11	F23(2) AND F29(1) => F32(1)
12	F23(2) AND F29(3) => F32(3)
13	F23(1) AND F29(3) => F32(1)
14	F2(1) AND F17(2) => F32(2)
15	F2(1) AND F17(1) => F32(1)
16	F2(2) AND F17(3) => F32(3)
17	F2(1) AND F17(3) => F32(2)
18	F17(2) AND F19(2) => F32(2)
19	F17(1) AND F19(1) => F32(1)
20	F17(2) AND F19(1) => F32(2)

Figure 7. Child table (2) the set of rules
图 7. 子表(2)的规则集

		Predicted			
		1	2	3	
Actual	1	2	1	0	0.666667
	2	0	3	0	1.0
	3	0	0	1	1.0
		1.0	0.75	1.0	0.857143

Figure 8. New case study rough set confusion matrix
图 8. 新例学习情况粗糙集混淆矩阵

机器学习的过程以测试子表(2)利用上述训练子表(1)生成的规则集进行学习。

%子表 2 机器学习过程如下：

%Note that the object indices below are 0-based.

Object 0： ok Actual = 2 (2)
Predicted = 2 (2)
Ranking = (0.666667) 2 (2) 2 rule(s)
(0.333333) 1 (1) 1 rule(s)

Object 1： ok Actual = 1 (1)
Predicted = 1 (1)
Ranking = (0.833333) 1 (1) 10 rule(s)
(0.166667) 2 (2) 2 rule(s)

Object 2： ok Actual = 2 (2)
Predicted = 2 (2)
Ranking = (0.666667) 2 (2) 8 rule(s)
(0.333333) 1 (1) 4 rule(s)

Object 3： ok Actual = 1 (1)
Predicted = 1 (1)
Ranking = (0.666667) 1 (1) 4 rule(s)
(0.333333) 2 (2) 2 rule(s)

Object 4： ok Actual = 3 (3)
Predicted = 3 (3)
Ranking = (0.764706) 3 (3) 6 rule(s)
(0.235294) 2 (2) 2 rule(s)

Object 5： ok Actual = 2 (2)
Predicted = 2 (2)
Ranking = (0.555556) 2 (2) 5 rule(s)
(0.444444) 3 (3) 3 rule(s)

Object 6： ERROR Actual = 1 (1)
Predicted = 2 (2)
Ranking = (0.666667) 2 (2) 4 rule(s)
(0.333333) 1 (1) 2 rule(s)

Confusion matrix：

	1	2	3	
1	2	1	0	66.66667%
2	0	3	0	100.0%
3	0	0	1	100.0%
	100.0%	75.0%	100.0%	85.71428%

估算结果：理解决策规则：如 F8(3) AND F22(1) AND F25(2) => F32(3)，根据实例工程的数据信息图 2 和离散图 3 对照，可以用如下语言来解释：当改路、改河、改渠填筑造价为 652,795，坡面植物防护造价为 0，护面墙造价为 86,236，则全生命周期造价为 21294604.27 元，其他规则依次类推。

从图示结果可知：7 个新例在根据训练样本子表(1)生成的学习规则学习时，其中第Ⅲ类(1 个测试样本)，第Ⅱ类(1 个测试样本)都能被正确判断(即图中数值为 100.0%)；但第Ⅰ类(3 个测试样本)的其中一个样本却被误判到第Ⅱ类；故 7 个等识样本有 6 个能识别，准确率(Actua1)达到 85.71428%。满足造价估算的精度要求。另外，为了提高识别率还可以通过学习新的规则或者通过多种学习算法增加断点(CUT)的角度提高识别率。

5. 结束语

粗糙集(RS)理论和机器学习目前理论研究比较多，但国内相关的实验系统还比较少，因此在研究基于 RS 理论的全生命周期造价时对数据的处理，一般还是用人工手算的方法比较多，很难在大型数据库系统的知识挖掘中完成相关知识的提取[8]，因此大大制约了 RS 理论在国内的研究和发展。本文利用 ROSETTA 实验系统，在 RS 理论的基础之上，经对工程量清单数据的处理，完成了新例的学习过程，准确率较高。算例分析表明，应用粗糙集机器学习解决全生命周期造价的投资估算是可行的。

参考文献 (References)

[1] 段晓晨, 张晋武, 李利军, 张健龙. 政府投资项目全面投资控制理论和方法研究[M]. 北京: 科学出版社, 2007: 12-30.
[2] 徐岳, 武同乐. 桥梁加固工程生命周期成本横向对比分析[J]. 长安大学学报(自然科学版), 2004, 24(3): 30-34.
[3] S. Yousefi, T. Hegazy, R. A. Capuruco, et al. System of multiple ANNs for online planning of numerous building improvements. Neurocomputing, 2008, 3(4): 346.
[4] 张勇. 粗糙集－神经网络智能系统在悬浮过程中的应用研究[D]. 大连: 大连理工大学, 2005.
[5] 孙士宝. 变精度粗糙集模型及其应用研究[D]. 四川: 西南交通大学, 2007.
[6] 交通部公路工程定额站. 公路工程工程量清单计量规则[M]. 长沙: 湖南省交通厅, 2005.
[7] 张云涛, 龚玲. 数据挖掘原理与技术[M]. 北京: 电子工业出版社, 2004.
[8] 程玉胜. Rosetta 实验系统在机器学习中的应用[J]. 安庆师范学院学报(自然科学版), 2005, 2: 69-72.