

运用粗糙集理论进行乡镇企业财务风险评价

吴 朗

(宜春学院 江西宜春 336000)

【摘要】 本文结合粗糙集理论的最新研究成果,找出乡镇企业不完备信息系统的内在规律并进行数据挖掘,然后根据收集到的数据进行知识约简和决策规则提取,以为企业的盈利水平和财务风险评价提供依据。

【关键词】 粗糙集理论 乡镇企业 数据挖掘 知识约简

一、乡镇企业不完备信息系统的数据挖掘方法

通常对乡镇企业不完备信息系统进行数据挖掘以前,应当先按照某种方法将其完备化,然后对完备化后的信息系统进行数据挖掘。但是,这种做法将会使数据失真,从而使获取的知识或提取的规则不可用。我们应当遵循分层递阶的原则,在本层次对那些容易采集的数据进行挖掘以获取所需的知识。如果在本层次无法获取所需的知识,就应进行深入观察,在下一个层次获取知识。这样就可以在不同的层次上分析同一问题,利用已有的知识逐步缩小问题求解的范围,直到得到最终结果。常见的处理不完备信息系统的方法有删除法、扩展法等,这些方法都不同程度地使获得的数据失真,并且未考虑属性获取的实时性、难易程度和成本等,从而影响最终决策的有用性。

我们根据图 2 进行横向分析可以得出,该高校个别收益性指标的值低于平均值,成长性指标、安全性指标的值均高于平均值,资源效率性指标的值也与平均值接近。从而,我们可以将该高校界定为开拓发展型高校。该高校今后应提高校办产业投资收益率,从而应加强学校的国有资产管理;应提高其他投资利润率,从而应充分利用学校人才、科研优势,增加学校收入,扩大资金来源;应提高资源利用率,从而应充分利用学校现有资源;应充分利用学校优势,扩大办学规模,在教学、科研、人才培养等方面再上一个新台阶,争取早日步入全面发展型高校之列。

我们根据图 2 进行纵向分析可以得出,与 2005 年相比,该高校 2006 年的教职工人均经费收入指标数值略有增大,教师人均科研经费指标数值有所降低。因此,该高校今后应充分发挥人才优势,加强科研合作,增加经费来源。该高校 2006 年的校办产业投资年收益率指标数值有所增大,但与平均值相比仍有一段差距。2006 年,该高校大部分成长性指标数值高于 2005 年,大部分安全性指标数值低于 2005 年。该高校 2005、2006 年两年大部分资源效率性指标的评价值接近平均值,这表明该高校发展潜力较大,按此趋势发展,该高校的收益性指标、资源效率性指标的数值会有进一步增大。

针对这些不完备信息系统处理方法存在的不足,本文从原始财务数据出发,直接对不完备信息系统进行数据挖掘,以提取最接近于原始数据的决策规则。数据挖掘的思路是:首先,从原始数据出发将属性分为完备层属性和不完备层属性,让不完备层属性出现在较高的层次上;然后,根据完备层属性建立首层决策子系统,确定目标规则的置信度,运用经典的粗糙集理论对其进行知识约简以得到满足置信度要求的规则;最后,根据首层决策子系统的知识约简情况和不完备层属性建立次层决策子系统,运用扩展的粗糙集理论对其进行规则推导。同时,在每个决策子系统中,根据推导出的规则构建模糊神经网络,以增强系统的决策能力。

二、粗糙集理论简介

设 U 是有限集,且 $U \neq \emptyset$;任何子集 $X \subseteq U$, X 是 U 中的

三、结语

基于雷达图对高校财务绩效进行分析,可以更加直观地反映出各项指标的评价值与平均值的差额以及高校未来的发展趋势,并预测指标值的未来走向。但要进一步分析各项指标的影响因素,则要运用因素分析法。如果要运用多个高校的多个财务绩效评价指标进行综合评价,则要确定各个指标的权重,并根据权重及指标值求出加权平均值。财务指标是评估高校发展状况的重要指标,但不是唯一指标,许多非财务指标如人力资源、科研能力、科研成果转化等方面的指标也能衡量高校的综合实力,我们也应对其加以考虑。

【注】 本文系河南省软科学研究项目“河南省高等学校财务绩效评价研究”(项目编号:082400450840)的阶段性研究成果。

主要参考文献

1. 杨周复,施建军.大学财务综合评价研究.北京:中国人民大学出版社,2002
2. 邵希娟,杜丽萍.财务分析中雷达图的阅读与绘制.中国管理信息化,2006;2
3. 张和生,张天武.高校财务绩效考核研究.财经论坛,2006;12

一个概念或范畴;R 是 U 中的一个等价关系;U/R 表示 R 的所有等价类;当 X 能代表某些 R 基本范畴的并时,称 X 为 R 可定义,否则称 X 为 R 不可定义。令:

$$\bar{R}X = U \setminus \{Y \in U/R \mid Y \subseteq X\}$$

$$RX = U \setminus \{Y \in U/R \mid Y \cap X \neq \emptyset\}$$

如果 $RX \neq \emptyset$ 且 $\bar{R}X \neq U$, 则称 X 为 R 粗糙可定义;如果 $RX = \emptyset$ 且 $\bar{R}X \neq U$, 则称 X 为 R 内不可定义;如果 $RX \neq \emptyset$ 且 $\bar{R}X = U$, 则称 X 为 R 外不可定义;如果 $RX = \emptyset$ 且 $\bar{R}X = U$, 则称 X 为 R 全不可定义。

由定义我们可以证得:集合 X 为 R 粗糙可定义(或为 R 全不可定义),当且仅当 $\sim X$ 为 R 粗糙可定义(或为 R 全不可定义);集合 X 为 R 外(内)不可定义,当且仅当 $\sim X$ 为 R 外(内)不可定义。

1. 不完备信息系统。设 $S=(U, AT)$ 是信息系统,其中:U 是对象的非空有限集合,AT 是属性的非空有限集合。对于每个 $a \in AT$,有 $a:U \rightarrow V_a$ 其中: V_a 称为 a 的值域。

每个属性子集 $A \subseteq AT$ 决定了一个不可区分关系 $IND(A)$:

$$IND(A) = \{(x, y) \in U \times U \mid \forall a \in A, a(x) = a(y)\}$$

$IND(A) (A \subseteq AT)$ 构成了 U 的划分,用 $U/IND(A)$ 来表示。

对于一个对象来说,一些属性值可能是缺省的。为了表明这种情况,通常对这些属性给定一个区分值(即空值)。

如果至少有一个属性 $a \in AT$ 使 V_a 含有空值,则称 S 为一个不完备信息系统;否则,它是完备的。我们用 nil 表示空值。

令 $A \subseteq AT$, 我们定义相似关系如下:

$$SIM(A) = \{(x, y) \in U \times U \mid \forall a \in A, a(x) = a(y) \text{ or } a(x) = \text{nil} \text{ or } a(y) = \text{nil}\}$$

我们可以证明 $SIM(A)$ 是一个相容关系:

$$SIM(A) = \bigcap_{a \in A} SIM(\{a\})$$

令 $S_A(x)$ 表示对象集 $\{y \in U \mid (x, y) \in SIM(A)\}$ 。对于 A 而言, $S_A(x)$ 是可能与 x 不可区分的对象的最大集合。

令 $D_A(x)$ 表示对象集 $\{y \in U \mid (x, y) \notin SIM(A)\}$ 。对于 A 而言, $D_A(x)$ 是可能与 x 可区分的对象的最大集合。

对任意 $x \in U, S_A(x) \cap D_A(x) = \emptyset$ 且 $S_A(x) \cup D_A(x) = U$ 。

令 $U/SIM(A) = \{S_A(x) \mid x \in U\}$ 表示分类。 $U/SIM(A)$ 中的任何元素称为相容类, $U/SIM(A)$ 中的相容类一般不构成 U 的划分而是构成 U 的覆盖, $\bigcup U/SIM(A) = U$ 。

2. 不可辨识关系与可辨识函数。设 $\Delta(x)$ 是不完备信息系统中对象 x 的区分函数,若:

$$\Delta(x) = \prod_{y \in U} a_{AT}(x, y)$$

其中: $(x, y) \in U$ 。

那么,每一个属性子集 $A \subseteq AT$, 对于每一个 $a \in AT$ 有 $a:U \rightarrow V_a$ 。如果满足 $\forall (a) \in A, Fa(x) = Fa(y)$, 则称对象 (x, y) 对于属性集合 A 而言是不可辨识的;否则,称对象 (x, y) 对于属性集合 A 而言是可辨识的。由 A 决定的不可辨识关系记为 $IND(A)$ 。

$$IND(A) = \{(x, y) \in U \times U \mid \forall (a) \in A, Fa(x) = Fa(y)\}$$

若出现信息丢失或信息不全的状况,属性可取空值 nil。在不完备信息系统中,无法明确辨识关系则视为相似。由 A 决定的相似关系记为 $SIM(A)$ 。

$$SIM(A) = \{(x, y) \in U \times U \mid \forall (a) \in A: f(x) = f(y) \text{ or } f(x) = \text{nil} \text{ or } f(y) = \text{nil}\}$$

$$S_A(x) = \{y \in U \mid (x, y) \in SIM(A)\}$$

研究人员提出了各种处理不完备信息系统的方法,但这里采用的方法不需对决策表进行任何转换,仅假设空值是属性中的某一个值。决策表可表示为 $DT=(U, AT \cup \{d\})$, 其中: d 为决策属性。定义函数:

$$\partial_A(x) = \{i \mid i = d(y), y \in S_A(x)\}$$

设 $a_A(x, y)$ 为属性 a 的集合,其中: $a \in A, (x, y) \notin SIM(\{a\})$ 。则有:

$$\Delta = \prod_{(x, y) \in U \times U} \sum a_{AT}(x, y)$$

其中: Δ 为 IS 的可辨识函数。

$$\Delta(x) = \prod_{y \in U} a_{AT}(x, y)$$

其中: $\Delta(x)$ 为对象 x 的可辨识函数。

$$\Delta^* = \prod_{(x, y) \in U \times \{\xi \in U \mid d(\xi) \notin \partial_{AT}(x)\}} \sum a_{AT}(x, y)$$

其中: Δ^* 为 DT 的可辨识函数。

3. 决策表、决策规则和知识约简。(不完备)决策表(DT)是一个(不完备)信息系统 $DT=(U, AT \cup \{d\})$ 。其中: d ($d \in AT$ 且 $\text{nil} \notin V_d$) 称为决策属性, AT 中的元素称为条件属性。

定义函数 $\partial_A: U \rightarrow p(V_d)$ 为:

$$\partial_A = \{i \mid i = d(y), y \in S_A(x)\}$$

称 ∂_A 为 DT 中的广义决策函数,其中: $p(V_d)$ 表示 V_d 的幂集。

如果对任意 $x \in U$ 有 $|\partial_{AT}(x)| = 1$, 则 DT 是协调(确定)的;否则,它是不协调(不确定)的。可以证明:关系 $IND(\partial_A), A \subseteq AT$ 构成了 U 的一个划分,且对于不完备 DT 有 $X \in U/IND(\partial_A) \Rightarrow \underline{A}X = X = \overline{A}X$ 不成立(尽管对于完备 DT 成立)。

任何决策表可以看做以下(广义)决策规则集:

$$\wedge(c, v) \rightarrow \vee(d, w)$$

其中: $c \in AT, v \in V_c, w \in V_d, \wedge(c, v)$ 称为规则的条件部分, $\vee(d, w)$ 称为规则的决策部分。

令 X 是具有性质 $\wedge(c, v)$ (其中: $c \in AT, v \in V_c$) 的对象集, Y 是具有性质 $\vee(d, w)$ (其中: $w \in V_d$) 的对象集。

在 DT 中,决策规则 $\wedge(c, v) \rightarrow \vee(d, w)$ 为真,当且仅当 $\overline{C}X \subseteq Y$ 。其中: C 是出现在规则的条件部分的所有属性构成的集合。

在 DT 中,决策规则 $r: \wedge(c, v) \rightarrow \vee(d, w)$ (其中: $c \in AT, v \in V_c, w \in V_d$) 是最优的,当且仅当该规则为真且由出现在 r 中的合取与析取的真子集构成的任何规则均为假。

三、应用案例

案例的有关资料见表 1, 其中: nil 代表空值。作为应用案

例,这里论域 $Q=\{1,2,3,4,5,6\}$ 分别代表六个企业,条件属性集 $A=\{P,M,S,X\}$ 分别代表企业性质、总产值增长率、净资产收益率、主营业务收入增长率。

表 1

	企业性质(P)	总产值增长率(M)	净资产收益率(S)	主营业务收入增长率(X)	企业盈利水平(D)
1	私企(B)	高(A)	高(A)	中(A)	一般(B)
2	国企(A)	nil	高(A)	中(A)	一般(B)
3	nil	nil	中(B)	高(B)	差(C)
4	私企(B)	nil	高(A)	高(B)	一般(B)
5	nil	nil	高(A)	高(B)	好(A)
6	国企(A)	高(A)	高(A)	nil	一般(B)

令 $Q/SIM(A)=\{S_A(1),S_A(2),\dots,S_A(6)\}$ 其中:

$S_A(1)=\{1\}$ $S_A(2)=\{2,6\}$ $S_A(3)=\{3\}$
 $S_A(4)=\{4,5\}$ $S_A(5)=\{4,5,6\}$ $S_A(6)=\{2,5,6\}$

令 $Q/IND=\{Y_1,Y_2,Y_3\}$ 其中:

$Y_1=\{1,2,4,6\}$ $Y_2=\{3\}$ $Y_3=\{5\}$

因而可得表 2:

表 2

N	1	2	3	4	5	6
∂_{AT}	{B}	{B}	{C}	{B,A}	{B,A}	{B,A}

从而可列出如下决策规则:

- $r_1:(P,B) \wedge (M,A) \wedge (S,A) \wedge (X,A) \rightarrow (D,B)$
- $r_2:(P,A) \wedge (M,nil) \wedge (S,A) \wedge (X,A) \rightarrow (D,B)$
- $r_3:(P,nil) \wedge (M,nil) \wedge (S,B) \wedge (X,B) \rightarrow (D,C)$
- $r_4:(P,B) \wedge (M,nil) \wedge (S,A) \wedge (X,B) \rightarrow (D,B) \vee (D,A)$
- $r_5:(P,nil) \wedge (M,nil) \wedge (S,A) \wedge (X,B) \rightarrow (D,B) \vee (D,A)$
- $r_6:(P,A) \wedge (M,A) \wedge (S,A) \wedge (X,nil) \rightarrow (D,B) \vee (D,A)$

我们通过条件属性约简以获取合适的规则,可得表 3 和表 4:

表 3

X/y	1	2	3	4	5	6
1		P	SX	X	X	P
2	P		SX	PX	X	
3	SX	SX		S	S	S
4	X	PX		S		P
5	X	X		S		
6	P			S	P	

这里有:

$$\Delta = P(S \vee X)X(P \vee X)S = PSX$$

$$\Delta(1) = P(S \vee X) = PX$$

$$\Delta(2) = P(S \vee X)X(P \vee X)X = PX$$

$$\Delta(3) = (S \vee X)S = S$$

$$\Delta(4) = X(P \vee X)S = PSX$$

$$\Delta(5) = SX$$

$$\Delta(6) = PS$$

表 4

X/y	1	2	3	4	5	6
1			SX		X	
2			SX		X	
3	SX	SX		S	S	S
4			S			
5			S			
6			S			

这里有:

$$\Delta^* = (S \vee X)XS = SX$$

$$\Delta^*(1) = (S \vee X)X = X$$

$$\Delta^*(2) = (S \vee X)X = X$$

$$\Delta^*(3) = (S \vee X)S = S$$

$$\Delta^*(4) = S$$

$$\Delta^*(5) = S$$

$$\Delta^*(6) = S$$

{s,x}是决策表 DT 的属性约简,{x}是对象 1 和 2 的相对约简。以上规则可简化为:

$R_1:(X,A) \rightarrow (D,B)$,即(主营业务收入增长率,中)→(企业盈利水平,一般)。

$R_2:(S,B) \rightarrow (D,C)$,即(净资产收益率,中)→(企业盈利水平,差)。

$R_3:(S,A) \rightarrow (D,B) \vee (D,A)$,即(净资产收益率,高)→(企业盈利水平,一般或好)。

以上案例运用粗糙集理论对乡镇企业的不完备信息系统的财务数据进行挖掘并进行知识约简,有效地划分了隐含知识。在对六个企业进行分析时,P(企业性质)、M(总产值增长率)不是决策要素,而 S(净资产收益率)、X(主营业务收入增长率)才是决策要素。通过知识约简,我们得到了三个最优风险评价规则,这些规则为企业的盈利水平和财务风险评价提供了依据。

主要参考文献

1. 张文修等.粗糙集理论与方法.北京:科学出版社,2004
2. 王国胤.Rough 集理论与知识获取.西安:西安交通大学出版社,2001
3. 刘清.Rough 集及 Rough 推理.北京:科学出版社,2005
4. 胡旺等.基于粗糙集理论不完备信息系统的数据挖掘.四川大学学报,2004;4
5. 李雄飞,李军.数据挖掘与知识发现.北京:高等教育出版社,2003
6. 乔斌,李玉榕,蒋静坪.粗糙集理论的分层递阶约简算法及其信息理论基础.控制理论与应用,2004;2