文章编号:1001 - 9081(2006)02 - 0368 - 03

# 应用决策树方法构建评价指标体系

陈 翔[1],刘军丽[2]
(1. 北京理工大学 管理与经济学院,北京 100081;2. 中国劳动关系学院 经济管理系,北京 100037)
(chenxiang@bit.edu.cn)

**摘 要**:在根据不同应用改进信息熵计算方法的基础上,提出了利用信息增益选择属性作为评价指标并得到其权重的方法。使用信息增益生成决策树,给出利用决策树计算指标评分细则的方法。最后,通过个人住房贷款信用风险评估体系的建立验证了这些方法的实用性。

**关键词**:信息熵;信息增益;决策树;评价指标体系
**中图分类号**: TP311 **文献标识码**:A

# Constructing evaluating indexes system with decision tree method

CHEN Xiang[1], LIU Jun-li[2]
(1. *School of Management and Economics, Beijing Institute of Technology, Beijing* 100081, *China*;
2. *Department of Economics & Management, China Institute of Industrial Relations, Beijing* 100037, *China*)

**Abstract:** According to different application, the computation formulas of information entropy were improved. These formulas to choose evaluating indexes and acquire their evaluation weight were applied. Based on decision tree that created by utilizing information gain, the detail rules of indexes rating were given. Finally, through constructing credit risk evaluating system of individual housing loan, the practicality of these methods was proved.

**Key words:** information entropy; information gain; decision tree; evaluating indexes system

在指标评价体系研究中,指标体系的构建是关键问题之一,构建合理的评价指标体系是科学评价的前提。因此,迫切需要对综合评价指标体系的构建方法进行研究。

造成评价指标多样性的主要原因是各个评价对象的性质的不一致,而且评价对象具有的数据属性及数据量呈现逐渐增多的趋势。

目前,建立指标体系的方法有很多,其中比较常见的有指数法和功效系数法,近几年,又出现了因子评价法、灰色关联度评价法、DEA 方法等新的评价方法[1]。这些方法都需要在对选择对象的每一个属性进行深入研究和计算后,才能得到适当的评价指标体系,而且不适应快速更新评价指标体系的需要。虽然也出现了一些应用数据挖掘构建评价指标体系的研究[2,3],但是这些研究没有给出评分的计算方法,也没有给出一个通用准则。决策树方法具有速度快、精度高、生成模式简单等优点,在数据挖掘的应用中得到许多软件公司和研究者的支持,有较多通用算法和研究成果[4]。本文在利用决策树方法的成果,特别是信息熵的计算和应用成果[5~7],以及在改进信息熵的计算方法,以适应不同的评价对象的基础上,得到快速选择最优指标集合及评分体系的方法,进一步利用信息增益生成决策树,给出计算二级指标评分体系的方法。最终,本文获得了一个具有普遍适用性的,全面的评价指标体系的构建方法。

## 1 信息熵计算方法的改进及决策树方法

决策树是一个类似于流程图的树结构,其中每个内部结点表示在一个属性熵的测试,每个分支代表一个测试输出,而每个树叶节点代表类或类分布。决策树通过把实例从根节点排列到某个叶子节点来分类实例,叶子节点即为实例所属的分类,树上每个节点说明了对实例的某个属性的测试,节点的每个后继分支对应于该属性的一个可能值。决策树算法包括 ID3、C4.5、SLIQ、SPRINT 等[4]。

其中,最核心的决策树学习算法为 ID3 算法,它是在所有可能的决策树空间中一种自顶向下、贪婪的搜索方法。ID3 算法的关键是确定属性表 $A$ 中可对训练例集 $E$ 进行的最佳分类的属性 $A$,即在树的每一个节点上确定一个候选属性,它的测试对训练实例的分类最有利。ID3 的搜索策略是爬山法,在构造决策树时从简单到复杂,用信息增益作为指导爬山法的评价函数。信息增益定义如下:

设 $S$ 是 $s$ 个数据样本的集合。假定类标号属性具有 $m$ 个不同类 $C_i(i=1,\cdots,m)$。设 $s_i$ 是类 $C_i$ 中的样本数。对一个给定的样本分类所需的期望信息由下式给出:

$$I(s_1,s_2,\cdots,s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

其中 $p_i$ 是任意样本属于 $C_i$ 的概率,并用 $s_i/s$ 估计。

设属性 $A$ 具有 $v$ 个不同值 $\{a_1,a_2,\cdots,a_v\}$。可以用属性 $A$ 将 $S$ 划分为 $v$ 个子集 $\{S_1,S_2,\cdots,S_v\}$。其中,包含 $S_j$ 中这样一些样本,它们在 $A$ 上具有值 $a_j$。如果 $A$ 选作测试属性(即最好的分裂属性),则这些子集对应于由包含集合 $S$ 的节点生长出来的分枝。设 $s_{ij}$ 是 $S_j$ 中类 $C_i$ 的样本数。根据由 $A$ 划分成子集的熵(entropy)或期望信息由下式给出:

$$E(A) = \sum_{i=1}^{v} \frac{s_{1j}+\cdots+s_{mj}}{s} I(s_{1j},\cdots,s_{mj}) \tag{2}$$

项 $\frac{s_{1j}+\cdots+s_{mj}}{s}$ 充当第 $j$ 个子集的权,并且等于子集(即 $A$

**收稿日期**:2005 - 08 - 26;**修订日期**:2005 - 11 - 04 **基金项目**:国家自然科学基金资助项目(70502021)

**作者简介**:陈翔(1976 - ),男,江西赣州人,讲师,博士,主要研究方向:Petri 网理论、数据挖掘技术和工作流技术; 刘军丽(1975 - ),女,河北邢台人,助教,硕士,主要研究方向:数据挖掘技术和统计分析方法.

值为 $a_j$) 中的样本个数除以 $S$ 中样本总数。熵值越小,子集划分的纯度就高。对于给定的子集 $S_j$:

$$I(s_{1j},s_{j2},\cdots,s_{mj}) = -\sum_{i=1}^{m} p_{ij}\log_2(p_{ij}) \tag{3}$$

其中,$p_{ij} = s_{ij}/s_j$ 是 $S_j$ 中的样本属于类 $C_i$ 的概率。

在 $A$ 上划分获得的信息增益为:

$$Gain(A) = I(s_1,s_2,\cdots,s_m) - E(A) \tag{4}$$

决策树算法就是计算每个属性的信息增益。将具有最高信息增益的属性选作给定集合 的测试属性,创建一个节点,并以该属性标记,对属性的每个值创建分枝,并据此划分样本。

决策树算法就是选择 $E(A)$ 最小的属性产生分支,这对左右分枝记录数相差不大的情况下是非常有效的[4]。如果左右分枝的记录数相差太远,用信息增益来判断可能得不到好的决策树。在评价目标的属性集中,由于应用领域的不同,可能会经常发生这种情况。为了解决这个问题,可以通过给熵加权的方式来解决。设 $A$ 为选择属性,有 $v$ 个属性值, 对应的权数为 $\omega_1,\omega_2,\cdots,\omega_v$,改进式(2) 为:

$$E(A) = \sum_{i=1}^{v} \omega_i \frac{s_{1j}+\cdots+s_{mj}}{s} I(s_{1j},\cdots,s_{mj}) \tag{5}$$

其中 $\omega_i$ 是指分支子集所占的比重,可以等于该分支记录数与整个记录集数量的比值。

在信息熵计算方法中,项 $\frac{s_{1j}+\cdots+s_{mj}}{s}$ 充当第 $j$ 个子集的权,取值较多的属性信息增益比较大。在评价指标选取过程中,当大多数属性取值较多、个别属性数取值较少,或者某一属性取值较多而这个属性在判断时没那么重要,利用式(2) 将掩盖取值少的属性的重要程度,也将增强取值多的属性的重要程度。

对于这种属性,可以引入优值法的概念对其进行改进[5]。给定 $0 < Q < 1$,$Q$ 称为用户选择优值法的参数,其大小由决策者根据先验知识或领域知识来确定。它是一个模糊的概念,通常指关于某一事务的先验知识,包括领域知识和专家建议,具体到评价指标选取则是指在除了用于评价指标选取的实例集之外的所有影响评价指标选择的因素。改进式(2) 为:

$$E(A) = \sum_{i=1}^{v} \left(\frac{s_{1j}+\cdots+s_{mj}}{s} + Q\right) I(s_{1j},\cdots,s_{mj}) \tag{6}$$

## 2　基于信息熵的评价指标选择及权重计算

设评价目标是由若干对象组成的有限集合,每个对象的含义由多个属性决定,每个定义四元组 $S = (U,A,V,f)$ 是一个指标系统,其中:

1) $U$:$U \neq \varnothing$,对象的非空有限集合;

2) $A$:代表综合评判的多种属性组成的集合称为因素集;

3) $V$:为多种评价构成的集合,称为评语集;

4) $f$:$U \times A \rightarrow V$ 是一个属性权重函数,它为每个对象的每个属性赋予一个权重值,即 $\forall a \in A, x \in U, f(x,a) \in V_a$。

目前选取指标的方法比较多,虽然这些方法的原理不尽相同,但都必须首先涉及到选择哪些指标来反映研究对象总体信息的问题。一般的,指标选择方法分为定性和定量两大类,并且要注意单个指标的意义和指标体系的内部结构。因为一方面既要所选的指标有代表性,能很好地反映研究对象某方面的特性;另一方面又要指标体系有全面性,能反映对象的全部信息。但若要满足全面性,势必要增加指标个数;但增加了指标个数,指标间相关程度可能性增大,反而影响了代表性。所以至今还没有一种方法能将代表性和全面性完美地综合起来,以及准确地衡量指标体系的有效程度。

目前常用的方法是选择一些后备指标后,利用统计中的一些方法,如相关系数法、条件广义最小方差法[1] 等,从中筛选出若干个有代表性的指标。虽然这些方法可以保证筛选出的指标相关性较低,但不能保证其确实完整地反映出研究对象的整体属性。例如相关系数法总是剔除与其他指标复相关系数最大的指标 ,但不能严格确定剔除多少指标后,剩下的指标能全面反映出研究对象的特性而相关性又最低。

在这里,我们利用信息增益解决这个问题。其大体思路是:指标体系的选择在遵循客观性、实用性、公开性等基本原则的基础上,主客观相结合。在数据处理过程中,不仅要按其原则进行数据处理,也要尽量根据经验补充可能会对决策产生重要影响的属性或计算参数。详细的指标选择步骤如下:

1) 选择 $n$ 个不同方面综合反映研究对象的整体特征,即根据先验知识得到评价子类,如对于个人住房贷款信用风险评估的应用,选择自然情况、购建房及还款能力情况、职业情况等 4 个大类作为评价子类。然后针对每个方面,选取若干候选属性。形成一个较全面的层次性很强的候选指标群。在个人住房贷款信用风险评估的应用中,选择拖欠月数、婚否、性别、年龄、职业、学历、房价收入比、保险情况等 18 个属性作为候选属性。

2) 根据不同的应用,选择不同的信息熵计算方法。在计算过程中,一般采用公式(4) 计算;如果评价目标左右分枝的记录数相差太远,可以考虑使用公式(5) 计算;如果候选属性中存在取值少而且重要的属性,可以考虑加入优值法参数,利用公式(6) 计算。

3) 信息增益的现实含义是从每个属性“获得的信息量”,代表该属性对结果类划分贡献的大小,因此,可以根据属性信息增益的大小获得该属性对区分评价对象产生影响的程度,信息增益小的属性对区分评价对象产生的影响小,反之影响程度大。给定一个信息增益阈值,将小于阈值的属性从候选属性删除,留下的候选属性就为其中的一个评价指标。

4) 对每个方面重复 2) ~3)。最后把各个方面筛选出来的指标集合起来,就是要找的属于评价指标的属性,设所获得的评价指标为 $A = \{a_1,a_2,\cdots,a_n\}$。

5) 各个评价指标的重要程度也能够通过信息增益得到,所以为了能够更好地反映决策属性和评价指标的关系,需要将各指标的重要性通过信息增益进行量化。建立一个函数,函数关系中的因变量为对决策属性的影响程度,影响程度越大,$y$ 值就越大,自变量为各评价指标的信息增益。为了符合指标计算的习惯,本文把 $Y$ 的值映射到 0 到 100 之间,同样,相应的各属性的取值也映射到 0 到 100 之间。即:

$$Y = INT\left(\frac{X}{\sum_{i=1}^{n} x_i} \times 100\right) \tag{7}$$

其中 $X = [x_1,x_2,\cdots,x_n]^T$ 为各评价指标的信息增益,$Y = [y_1, y_2, \cdots, y_n]^T$ 为这些指标对应的评价权重。

## 3　基于决策树确定指标评分细则

上一小节得到的指标评价权重只能描述某个属性在决策中的重要程度,不能反映该属性各个取值的重要程度。在获得各个属性的评价权重之后,可以将对象的非空有限集合 $U$ 所构成的数据集用于构建指标的评分细则,其基本思路是:根据

已知的样本与评价目标的对应关系,运用决策树发现评价集 $V$ 与评价指标之间的关系,使得能够通过对评价属性的具体观察值,对评价目标的真实情况进行预测。详细的步骤如下:

1) 采用层次分析法[1] 等方法确定各评价取值相对于评价集的权重分配,即确定评价集中多种评价的相对重要程度。相对重要程度也可以通过领域专家的调查问卷,通过统计方法获得(由于这些方法不是本文重点,在此不做详细叙述),设所得到的评价集 $V$ 每一个取值 $V_m$ 的相对重要程度为 $W = \{w_1, w_2, \cdots, w_m\}$;

2) 运用决策树方法[4] 构造评价对象数据集的决策树(由于决策树方法也不是本文重点,在此不做详细叙述),在决策树中可以获得评价指标不同取值下对应各种评价的对象个数。如果指标在某个取值下,含有较多评价为重要程度高的评价对象,那么该取值就相对重要,所占的评分相对就较高,反之评分相对就较低。在这种思想的指导下,首先构造指标每一个取值的"评分细则度量值"$z_j$,计算公式如下:

$$z_j = \frac{W \times R_j}{\sum_{i=1}^{m} r_{jm}} \tag{8}$$

其中 $R_j = \{r_{j1}, r_{j2}, \cdots, r_{jm}\}^T$,对应某个指标的取值 $j$ 中含有评价取值为 $V_m$ 的对象的个数;

3) 根据"评分细则度量值" 可以构造公式最后确定每个指标各个取值的分数值 $\lambda_j$:

$$\lambda_j = INT\left(\frac{z_j}{\max_{j=1}^{n}(z_j)} \times y_j\right) \tag{9}$$

其中 $n$ 为指标取值的个数,$y_i$ 为评价指标的权重,$INT$ 表示取整函数。式(9) 表明"评分细则度量值" 最高的指标取值,它的评分值取该指标的评价权重值,其他评价取值按该值的"评分细则度量值" 与最高的"评分细则度量值" 之比值计算评分值;

4) 重复 2) ~3),直至所有评价指标的所有取值的评分细则计算完毕。

通过评价指标的获得,在确定每个评价指标权重的基础上,获得每个指标所有取值的评分细则后,就建立了一个比较完整的指标评价体系。

## 4 案例

以个人住房贷款信用风险评估体系的建立过程来说明前面的方法。该案例采用了某银行 2003 年以前的个人住房贷款数据,案例的评价目标是新贷款用户的风险状况,其评语集为{高风险贷款, 中等风险贷款, 较低风险贷款,低风险贷款}。根据先验知识,选择自然情况、购建房及还款能力情况、职业情况等 4 个大类作为评价子类。针对每个评价子类,在个人住房贷款数据中选择拖欠月数、婚否、性别、年龄、职业、学历、房价收入比、保险情况等 18 个属性作为候选属性,其中身份证号为主关键字,不参与运算。

该案例中左右分支结点数量不大,候选属性的数据量也比较平均,因此,选择式 2 计算信息熵,再利用式 4 得到这些属性的信息增益分别是:

| | |
|---|---|
| Gain(还贷收入比) = 0.2525 | Gain(房价收入比) = 0.1616 |
| Gain(职业) = 0.0984 | Gain(岗位) = 0.0630 |
| Gain(学历) = 0.0561 | Gain(年龄) = 0.0422 |
| Gain(保险) = 0.0283 | Gain(性别) = 0.0012 |
| Gain(婚否) = 0.0008 | …… |

设置的阈值为 0.01,所以"性别"和"婚否"等 10 个属性的信息增益太小,这 10 个属性可以删除,留下的 7 个属性即为评价指标。根据式 7 解得这了:些指标对应的评价权重 $Y = [36, 23, 14, 9, 8, 6, 4]^T$。

在案例中,采用统计方法得到的评价集的相对重要程度为 $W = \{0.4, 0.15, 0.02, 0\}$,对应于不同风险评价违约的概率。如果风险越高,则其"风险度量值" 就越高。通过在决策树中获得属性 每个取值 $j$ 中"高风险贷款"、"中等风险贷款"、"较低风险贷款" 和"低风险贷款" 出现人数的数量 $r_{j1}$、$r_{j2}$、$r_{j3}$ 和 $r_{j4}$,由下式求得指标各个取值的"风险度量值":

$$z_j = (0.4 \times r_{j1} + 0.15 \times r_{j2} + 0.02 \times r_{j3} + 0 \times r_{j4}) / (r_{j1} + r_{j2} + r_{j3} + r_{j4}) \tag{10}$$

通过上述计算,可以得到各个指标的评分细则,例如可以计算得到还贷收入比指标的评分细则如下:

| | | | |
|---|---|---|---|
| 还贷收入比 | 31: 低还贷收入比 | (0, 0.3) | 36 |
| | 32: 较低还贷收入比 | (0.3, 0.4] | 28 |
| | 33: 中还贷收入比 | (0.4, 0.6) | 14 |
| | 34: 较高还贷收入比 | (0.6, 1.0) | 5 |

通过计算年龄、职业等其他 6 个指标取值的评分细则,就可以得到个人住房贷款信用风险评估体系。根据该体系可以很方便地根据每一个个人住房贷款申请人填写的申请表得到一个确切的、可以度量的值,对于分值高的客户可以快速审察,对于分值低的用户需要综合考虑其他因素再考虑是否给予贷款。这在很大程度上可以帮助信贷人员进行审核,为管理部门快而有效地进行决策提供支持。

## 5 结语

构建合理的指标体系是对评价对象科学评价的重要前提,文中在分析评价指标与评价对象属性之间关系的基础上,应用决策树技术对指标体系构建进行了研究,提出了构建评价指标体系的方法。根据信息增益和信息熵的计算,建立指标属性的选取原则和权重分析方法,利用决策树给出了指标体系评分细则的方法。该方法是一种通用的指标体系构建方法,可以用于为大多数评价对象建立综合评价方法。应该指出的是,随着对评价对象认识的加深和与之相关的核算方法的变化,针对同一对象的评价指标体系也会随之变化,因此应根据需要更新评价指标体系。此外,对一些比较特殊的评价对象,还需要构造与之适应的信息熵计算方法。

**参考文献:**


[1] 胡永宏, 贺思辉. 综合评价方法 [M]. 北京: 科学出版社, 2000.

[2] SHENOY PP. A comparison of graphical techniques for decision analysis [J]. European Journal of Operational Research, 1994, 78 (1): 1 - 21.

[3] 赵晓冬, 郑涛. 基于 FUZZY-AHP 评价方法的个人信用等级评价模型指标体系 [J]. 数量经济技术经济研究, 2003, (6): 97 - 100.

[4] HAN JW, KAMBER M. 数据挖掘: 概念与技术 [M]. 范明, 孟小峰, 等译. 北京: 机械工业出版社, 2001.

[5] QUINLAN JR. Improved use of continous attributes in C4.5 [J]. Journal of Artificial Intelligence Research, 1996, 24(4): 77 - 90.

[6] 张维东, 张凯, 董青, 等. 利用决策树进行数据挖掘中的信息熵计算[J]. 计算机工程, 2001, 27(3): 87 - 89.

[7] CARDIE C. Using decision trees to improve case - based learning [A]. In: Proceedings of the Tenth International Conference on Machine Learning [C]. Morgan Kaufmann Publishers, Inc. 1993. 25 - 32.