基于极大熵聚类的工程项目风险预警模型

唐葆君 刘小龙 邱菀华

(北京航空航天大学 经济管理学院，北京 100191)

摘 要：该方法针对当前大部分工程风险预警模型只能报警，不能预测的现状，提出了基于熵最优化的工程项目风险预警方法.利用判别熵最小化选取项目风险预警指标值，通过特征选取找出那些最有效的特征，研究出一种新的聚类算法——极大熵聚类算法，极大熵聚类算法是以概率为比例将任一指标向量分配给所有码向量，而不是仅仅只分配给与之最近的码向量，该算法是C-均值算法的一种推广.最后用实例验证该模型，用此算法对预测结果进行分类，判断项目的风险状态.结果表明这种方法估计工程项目风险快捷有效，与实际情况基本一致，可以应用于工程分析.

关 键 词：极大熵聚类；判别熵；特征提取；风险预警

中图分类号：F 540

文献标识码：A **文 章 编 号**：1001-5965(2008)07-0812-04

Early-warning of engineering-project risk model research based on maximum entropy clustering

Tang Baojun Liu Xiaolong Qiu Wanhua

(School of Economics and Management, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Abstract: Due to most of current early-warning of engineering risk only give a alarm, but cannot forecast, engineering-project risk early-warning based on entropy optimal model was proposed. Firstly, minimum J-divergence entropy was applied to extract the risk early-warning index and to find out the most effective feature by feature extraction. Then the calculating result was classified to judge state of project with a new clustering algorithm-maximum entropy clustering algorithm. Maximum entropy clustering algorithm allocated index vectors to all of the code vectors rather than the nearest code vector with a ratio of possibility. The algorithm is an improved c-means algorithm. Finally, the case was verified to, the results of forecast were classified by the algorithm to estimate the project's venture. The experiment results show that the improved algorithm can use to predict the project risk quickly and effectively at engineering analyses. The analytical results are basically identical with the actual situation.

Key words: maximum entropy clustering; J-divergence entropy; feature extraction; risk early-warning

工程项目风险预警是指根据工程实际特点，针对不同建设阶段，研究评价各风险因素，确定风险级别，建立预警系统，采取风险防范和控制措施，以降低风险损失，保证工程建设顺利进行的过程.但由于工程项目十分复杂，在风险预警过程中，人们无法全面、准确地辨别出项目的真实状态，只能依靠一些指标(特征)来判断其所属类别，显然这些指标的选择是很重要的，它强烈影响到分类器的设计及其性能[1].但目前项目指标选取和评价主要以专家咨询和评分为主.因此，解决特征选取问题是项目风险预警模型设计中的一个关键问题，本文利用判别熵最小化和专家评价相结

收稿日期：2007-06-28

作者简介：唐葆君(1972—)，女，江西泰和人，博士生，tang_baojun@sohu.com.

合的方法完成预警模型中的特征选取,在风险等级分类器的设计中借助极大熵原理构造了一致逼近目标函数的一簇可微的熵函数,由此利用最优化理论导出了一种新的聚类算法—极大熵聚类算法.该算法是 C-均值算法的一种推广.

1　项目风险预警模型设计

1.1　基于判别熵最小化的特征选取

描述工程项目经营管理状况的指标涉及到技术开发风险、生产风险、市场风险、资金风险、金融风险、技术管理风险、社会环境风险等各个方面.特征选取基本任务是从许多特征中找出那些最有效的特征.

作为不确定性的一种度量的表达式 shannon 熵[1],这样一种概念也可以用来作为某个概率分布密度 $p(\boldsymbol{x}_i)$ 偏离给定标准分布 $\omega(\boldsymbol{x}_i)$ 的程度的度量,把它叫做相对熵,即

$$V(p,\omega)=-\sum p(\boldsymbol{x}_i)\log[p(\boldsymbol{x}_i)/\omega(\boldsymbol{x}_i)]\leqslant 0 \tag{1}$$

求和应在该特征所有可能的取值上进行.相对熵越小,这两类概率分布的差别就越大,但两类概率分布完全相同时,相对熵达最大值(等于零).

根据相对熵的特点,可以定义判别熵 $W(p,q)$ 来表征两类分布 $p(\boldsymbol{x}_i)$ 和 $q(\boldsymbol{x}_i)$ 的差别大小.

$$\begin{aligned}w(p,q)=&V(p,q)+V(q,p)=\\&-\sum p(\boldsymbol{x}_i)\log p(\boldsymbol{x}_i)-\sum q(\boldsymbol{x}_i)\log q(\boldsymbol{x}_i)+\\&\sum p(\boldsymbol{x}_i)\log q(\boldsymbol{x}_i)+\sum q(\boldsymbol{x}_i)\log p(\boldsymbol{x}_i)\leqslant 0\end{aligned} \tag{2}$$

在多种情况下,可以用 $\sum_i\sum_j W(p^{(i)},q^{(j)})$ 来表示各类分布之间的分离程度.i,j 代表类别号.

对于特征提取来说,在给定维数 d 的条件下,应该求得这样 d 个指标,它使上述判别熵最小.

为了计算方便起见,可以用下列函数

$$U(p,q)=-\sum_i(p_i-q_i)^2\leqslant 0 \tag{3}$$

来代替 $W(p,q)$,而不影响选取 d 个最优指标的结果.

在不对概率分布作估计的情况下,可以用经过归一化处理的样本指标值来代替上式中的概率分布:

$$\left.\begin{aligned}p_i^{(1)}&=\frac{1}{N_1}\sum_{k=1}^{N_1}(x_{ki}^{(1)})^2\\\sum_{i=1}^{D}(x_{ki}^{(1)})^2&=1\end{aligned}\right\} \tag{4}$$

式中,k 是第一类样本集中的样本号;N_1 是第一类的样本总数;i 是特征号.由于 $\sum_{i=1}^{D}p_i=1$,所以这样做是合理的.同理计算 q_i.

1.2　基于极大熵聚类的分类设计

若按判别熵最小原则,即依据式(3),式(4),在 n 个项目样本组成的样本集中,选定样本的 k 个指标,指标矩阵可表示为

$$\boldsymbol{X}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&\vdots&&\vdots\\x_{k1}&x_{k2}&\cdots&x_{kn}\end{bmatrix} \tag{5}$$

式中,x_{ij} 为样本 j 指标 i 的指标值,$i=1,2,\cdots,k$,$j=1,2,\cdots,n$.亦可将 $\boldsymbol{X}$ 表示成一个指标向量集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k\}\subset\mathbf{R}^n$.若将项目风险状态等级总数记为 c,即 c 个评语等级,考虑某种相似性度量,则可将 $\boldsymbol{X}$ 聚合成 c 个分离开的子集 $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_c$.每个子集表示一类,分别包括 $n_1,n_2,\cdots,n_c$ 个指标向量,设 $\boldsymbol{X}_j=\{x_1^{(j)},x_2^{(j)},\cdots,x_{n_j}^{(j)}\}$.为了衡量聚类的质量,常采用误差平方和 J 为目标函数,即

$$J(\boldsymbol{X},\boldsymbol{V})=\sum_{j=1}^{c}\sum_{\boldsymbol{x}_k\in\boldsymbol{X}_j}p(\boldsymbol{x}_k)\parallel\boldsymbol{x}_k-\boldsymbol{v}_j\parallel^2 \tag{6}$$

式中,$p(\boldsymbol{x}_k)$ 是抽样概率;$\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\}\subset\mathbf{R}^n$ 是码本;而 $\boldsymbol{v}_j(j=1,2,\cdots,c)$ 称为码向量;它被定义为第 j 类 $\boldsymbol{X}_j$ 的均值,即 $\boldsymbol{v}_j=\frac{1}{n_j}\sum_{k=1}^{n}x_k^{(j)}$;指标向量 $\boldsymbol{x}_k\in\boldsymbol{X}_j$ 由最近邻原理定义,即当 $\parallel\boldsymbol{x}_k-\boldsymbol{v}_j\parallel^2=\min_l\parallel\boldsymbol{x}_k-\boldsymbol{v}_l\parallel^2$,$\boldsymbol{x}_k\in\boldsymbol{X}_j$.C- 均值算法就是由最近邻原理和 $\boldsymbol{v}_j=\frac{1}{n_j}\sum_{k=1}^{n}x_k^{(j)}$ 定义的,但它是一个启发过程,而不是一个最优化过程,这是由于函数 J 是不可微的,于是无约束最优化的梯度方法不能够直接应用.这类算法最大的问题是算法训练没有一个终止准则,算法最后结果严重依赖码向量的初始值.

从最优化理论的角度看,求式(6)的最小问题其实是个不可微优化问题.人们常用一簇可微函数逼近目标函数来处理该问题.比如,文献[2]借助极大熵原理构造了一致逼近目标函数的熵函数,本文根据这一思想,研究了一种极大熵聚类算法进行分类.

对于指标向量 $\boldsymbol{x}$,定义函数 $\boldsymbol{f}_{\boldsymbol{x}}:\mathbf{R}^{cn}\to\mathbf{R}^c$,$\boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{V})=\boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c)=(-\parallel\boldsymbol{x}-\boldsymbol{v}_1\parallel^2,-\parallel\boldsymbol{x}-\boldsymbol{v}_2\parallel^2,\cdots,-\parallel\boldsymbol{x}-\boldsymbol{v}_c\parallel^2)^{\mathrm{T}}$,则 $\parallel\boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{V})\parallel_\infty=\max_j(-\parallel\boldsymbol{x}-\boldsymbol{v}_i\parallel^2)=$

$-\min_j(\|\boldsymbol{x}-\boldsymbol{v}_j\|)^2$. 于是式(6)可以改写为

$$J(\boldsymbol{X},\boldsymbol{V})=\sum_{k=1}^{K}p(\boldsymbol{x}_k)(-\|f_{\boldsymbol{x}_k}(\boldsymbol{V})\|_\infty) \tag{7}$$

下面引入一种新的衡量聚类质量的目标函数,

$$J_c=J_c(\boldsymbol{X},\boldsymbol{V})=$$

$$\sum_{j=1}^{c}\sum_{\boldsymbol{x}_k\in\boldsymbol{X}_j}p(\boldsymbol{x}_k)p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)\|\boldsymbol{x}_k-\boldsymbol{v}_j\|^2=$$

$$\sum_{i=1}^{c}\sum_{\boldsymbol{x}_k\in\boldsymbol{X}_j}p(\boldsymbol{x}_k,\boldsymbol{v}_j)\|\boldsymbol{x}_k-\boldsymbol{v}_j\|^2=$$

$$\sum_{k=1}^{K}p(\boldsymbol{x}_k)J_c(\boldsymbol{V}\mid\boldsymbol{x}_k) \tag{8}$$

式中,$p(\boldsymbol{x}_k,\boldsymbol{v}_j)$是 $\boldsymbol{x}_k$ 和 $\boldsymbol{v}_j$ 的联合概率;$p(\boldsymbol{v}_j,\boldsymbol{x}_k)$是条件概率;$J_c(\boldsymbol{V}|\boldsymbol{x}_k)$定义为

$$J_c(\boldsymbol{V}\mid\boldsymbol{x}_k)=J_c(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\mid\boldsymbol{x}_k)=$$

$$\sum_{j=k}^{c}p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)\|\boldsymbol{x}_k-\boldsymbol{v}_j\|^2 \tag{9}$$

当完全分配一个指标向量给与之最近的码向量,即条件概率由下式定义时,

$$p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)=\begin{cases}1, & \boldsymbol{x}_k\in\boldsymbol{X}_j\\ 0, & \text{其他}\end{cases} \tag{10}$$

式(8)退化为式(6). 此时关于自由参数$\{\boldsymbol{v}_j,p(\boldsymbol{v}_j|\boldsymbol{x}_k)\}(k=1,2,\cdots,K;j=1,2,\cdots,c)$求式(8)定义的 J_c 最小,立即可以产生一个硬聚类解. 然而可以把这个最优化问题考虑为去寻找一个分布,在满足一定程度随机性下它最小化目标函数 J_c. 自然,随机程度可以用 $\boldsymbol{X}$ 和 $\boldsymbol{V}$ 的 Shannon 联合熵来度量,即

$$H(\boldsymbol{X},\boldsymbol{V})=-\sum_{k=1}^{K}\sum_{j=1}^{c}p(\boldsymbol{x}_k,\boldsymbol{v}_j)\ln p(\boldsymbol{x}_k,\boldsymbol{v}_j) \tag{11}$$

于是这个最优化问题可以简单地变为 Lagrange 最小化问题:

$$L(\boldsymbol{X},\boldsymbol{V})=J_c(\boldsymbol{X},\boldsymbol{V})-TH(\boldsymbol{X},\boldsymbol{V}) \tag{12}$$

T 是 Lagrange 乘子. 很明显,对大的 T,主要是试图最大熵,随着 T 的降低,以熵换取失真的减少,当 T 趋于零,最小 J_c 直接获得一个非随机解.

进一步分析由式(12)定义的 Lagrange 函数 L,首先注意到可以分解联合熵:

$$H(\boldsymbol{X},\boldsymbol{V})=-\sum_{k=1}^{K}\sum_{j=1}^{c}p(\boldsymbol{x}_k,\boldsymbol{v}_j)\ln p(\boldsymbol{x}_k,\boldsymbol{v}_j)=$$

$$-\sum_{k=1}^{K}\sum_{j=1}^{c}p(\boldsymbol{x}_k)p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)\ln(p(\boldsymbol{x}_k)p(\boldsymbol{v}_j\mid\boldsymbol{x}_k))=$$

$$-\sum_{k=1}^{K}\sum_{j=1}^{c}p(\boldsymbol{x}_k)p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)(\ln p(\boldsymbol{x}_k)+$$

$$\ln p(\boldsymbol{v}_j\mid\boldsymbol{x}_k))=-\sum_{k=1}^{K}p(\boldsymbol{x}_k)\ln p(\boldsymbol{x}_k)-$$

$$\sum_{k=1}^{K}p(\boldsymbol{x}_k)\sum_{j=1}^{c}p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)\ln p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)=$$

$$H(\boldsymbol{X})+\sum_{k=1}^{K}p(\boldsymbol{x}_k)H(\boldsymbol{V}\mid\boldsymbol{x}_k)=$$

$$H(\boldsymbol{X})+H(\boldsymbol{V}\mid\boldsymbol{X}) \tag{13}$$

其中

$$H(\boldsymbol{X})=-\sum_{k=1}^{K}p(\boldsymbol{x}_k)\ln p(\boldsymbol{x}_k) \tag{14}$$

$$H(\boldsymbol{V}\mid\boldsymbol{x}_k)=H(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\mid\boldsymbol{x}_k)=$$

$$-\sum_{j=1}^{c}p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)\ln p(\boldsymbol{v}_j\mid\boldsymbol{x}_k) \tag{15}$$

$$H(\boldsymbol{V}\mid\boldsymbol{X})=$$

$$-\sum_{k=1}^{K}p(\boldsymbol{x}_k)\sum_{j=1}^{c}p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)\ln p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)=$$

$$\sum_{k=1}^{K}p(\boldsymbol{x}_k)H(\boldsymbol{V}\mid\boldsymbol{x}_k) \tag{16}$$

由于 $H(\boldsymbol{X})$是信源熵,它独立于聚类,因而可以从函数 L 中抽取常数 $H(\boldsymbol{X})$,而主要集中于条件熵 $H(\boldsymbol{V}|\boldsymbol{X})$. 另一方面,由式(6)定义的目标函数 J 总是非负的,自然也希望逼近它的函数 L 同样非负,但由于$-H(\boldsymbol{V}|\boldsymbol{X})$是负的,因而这个条件不能够得到保证. 然而知道 $H(\boldsymbol{V}\mid\boldsymbol{X})=\sum_{k=1}^{K}p(\boldsymbol{x}_k)\cdot H(\boldsymbol{V}|\boldsymbol{x}_k)\leqslant\sum_{k=1}^{K}p(\boldsymbol{x}_k)\ln C=\ln C$,并且 $\min(-H(\boldsymbol{V}|\boldsymbol{X}))$同 $\min(\ln C-H(\boldsymbol{V}|\boldsymbol{X}))$是等价的. 于是可以把最小化问题式(12)变为

$$\min\{L_{\mathrm{T}}(\boldsymbol{X},\boldsymbol{V})=J_c(\boldsymbol{X},\boldsymbol{V})+$$

$$T(\ln C-H(\boldsymbol{V}\mid\boldsymbol{X}))=$$

$$\sum_{k=1}^{K}p(\boldsymbol{x}_k)L_{\mathrm{T}}(\boldsymbol{V}\mid\boldsymbol{x}_k)\} \tag{17}$$

$$L_{\mathrm{T}}(\boldsymbol{V}\mid\boldsymbol{x}_k)=J_c(\boldsymbol{V}\mid\boldsymbol{x}_k)+T(\ln C-H(\boldsymbol{V}\mid\boldsymbol{x}_k)) \tag{18}$$

关于 $p(\boldsymbol{v}_j,\boldsymbol{x}_k)$直接最小化 $L_{\mathrm{T}}(\boldsymbol{X},\boldsymbol{V})$,可以得到 $p(\boldsymbol{v}_j|\boldsymbol{x}_k)$满足 Gibbs 分布,即

$$p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)=\frac{\exp[-\|\boldsymbol{x}_k-\boldsymbol{v}_j\|^2/T]}{Z_{\boldsymbol{x}_k}} \tag{19}$$

正则参数 $Z_{\boldsymbol{x}_k}$为

$$Z_{\boldsymbol{x}_k}=\sum_{j=1}^{C}\exp[-\|\boldsymbol{x}_k-\boldsymbol{v}_j\|^2/T] \tag{20}$$

把式(19)代入到式(17),就可以获得函数 $L_{\mathrm{T}}(\boldsymbol{X},\boldsymbol{V})$相应的最小形式 $L_{\mathrm{T}}^*(\boldsymbol{X},\boldsymbol{V})$:

$$L_{\mathrm{T}}^*(\boldsymbol{X},\boldsymbol{V})=\min_{p(\boldsymbol{v}_j|\boldsymbol{x}_k)}L_{\mathrm{T}}(\boldsymbol{X},\boldsymbol{V})=T\sum_{k=1}^{K}p(\boldsymbol{x}_k)\cdot$$

$$\left[\ln C-\ln\sum_{t=1}^{c}\exp[-\|\boldsymbol{x}_k-\boldsymbol{v}_j\|^2/T]\right]=$$

$$\sum_{k=1}^{K}p(\boldsymbol{x}_k)L_{\mathrm{T}}^*(\boldsymbol{V}\mid\boldsymbol{x}_k) \tag{21}$$

其中,$L_{\mathrm{T}}^*(\boldsymbol{X},\boldsymbol{V})$称为熵函数(或凝聚函数).

$$L_{\mathrm{T}}^{*}(\boldsymbol{V}\mid\boldsymbol{x}_k)=$$
$$-T\ln\sum_{j=1}^{C}\exp\left[-\frac{\|\boldsymbol{x}_k-\boldsymbol{v}_j\|^2}{T}\right]+T\ln C \quad (22)$$

关于码向量 $\boldsymbol{v}_j$ 最小化 $L_{\mathrm{T}}^{*}(\boldsymbol{X},\boldsymbol{V})$,设置它的梯度为零,经过简单计算得到:

$$\sum_{k=1}^{K}p(\boldsymbol{x}_k)p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)(\boldsymbol{x}_k-\boldsymbol{v}_j)=0 \quad (23)$$

或

$$\sum_{k=1}^{K}p(\boldsymbol{x}_k,\boldsymbol{v}_j)(\boldsymbol{x}_k-\boldsymbol{v}_j)=0 \quad (24)$$

所以

$$\boldsymbol{v}_j=\frac{\sum_{k=1}^{K}p(\boldsymbol{x}_k,\boldsymbol{v}_j)\boldsymbol{x}_k}{\sum_{k=1}^{K}p(\boldsymbol{x}_k\mid\boldsymbol{v}_j)}=\frac{\sum_{k=1}^{K}p(\boldsymbol{v}_j)p(\boldsymbol{x}_k\mid\boldsymbol{v}_j)\boldsymbol{x}_k}{p(\boldsymbol{v}_j)}=$$
$$\sum_{k=1}^{K}p(\boldsymbol{x}_k\mid\boldsymbol{v}_j)\boldsymbol{x}_k \qquad j=1,2,\cdots,c \quad (25)$$

在实际应用中,一般都假设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k\}\subset\mathbf{R}^N$ 中的指标向量是相互独立的,即 $p(\boldsymbol{x}_k)=1/K$ 时,此时由式(23)可得

$$\boldsymbol{v}_j=\frac{\sum_{k=1}^{K}p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)\boldsymbol{x}_k}{\sum_{k=1}^{K}p(\boldsymbol{v}_j\mid\boldsymbol{x}_k)} \qquad j=1,2,\cdots,c \quad (26)$$

综上所述,由于该聚类算法起源于极大熵函数,称之为极大熵聚类算法.

可以看出,当 $T\to 0$ 时极大熵聚类算法就是 C-均值算法,由此可以说极大熵聚类算法是 C-均值算法的一种推广格式;而当 $T\to\infty$ 时,由于所有的码向量的值相等,因而实际上是把所有的指标向量聚成一类.由式(25)或式(26)可知,算法对于任一指标向量是以概率为比例分配给所有码向量,而不是仅仅只完全分配给与之最近的码向量,它在一定程度上能克服标准 C-聚类对初始码向量选取敏感的问题.而且算法对具有噪声干扰的指标向量进行聚类时具有较强的鲁棒性.

2　实例研究

以某高速公路路面结构设计和道路工程质量检验为例,对公路工程质量进行风险预警[3].

2.1　特征提取

根据公路工程质量评定国家标准,将评定要素取为特征集,利用专家评价法从中选择 10 个特征,对每一个要评判的具体工程,其评定结果可分为 2 类:优,劣.通过德尔斐法选择 12 条优质和 12 条劣质高速公路的数据作为训练样本.现要将 10 个指标值通过特征提取后得到 4 个指标,用 p_i 表示第 1 类(劣质)样本的第 i 个指标值的概率分布,q_i 表示第 2 类(优质)样本的第 i 个指标值的概率分布,通过式(6)计算得出 p_i 和 q_i 的值.为保证式(5)成立,只需找出使 $-(p_i-q_i)^2$ 最小的 4 个指标即可.从而得到特征提取后的 4 个指标为:路基路面,安全舒适,桥梁涵洞,排水防护.

2.2　辨别测试样本类别

将某高速公路路面结构设计和道路工程质量检验作为测试样本,取出其 4 个指标值.根据极大熵聚类算法的步骤:定义危机等级 $c=2$,Lagrange 乘子 $T(t)=1/\beta(t)$,其中 $\beta(t)$ 的选取规则为 $\beta(t)=\beta_0\alpha^{(t)}$,$t$ 代表当前迭代次数,$\beta_0=0.5$,$\alpha=1.1$.表 1 给出了实证中采用的码向量的初始值,该初始值的选取方法见文献[4].

表 1　码向量的初始值

初始值	路基路面	安全舒适	桥梁涵洞	排水防护
第 1 类初始值	0.1	0.3	0.5	0.3
第 3 类初始值	0.2	0.4	0.9	0.9

利用初始值迭代计算可知,该工程质量为第 1 类的条件概率值为 0.153,为第 2 类的条件概率值为 0.847.故该工程质量可以评为优.

3　结 束 语

研究分析表明,运用熵方法可以有效、准确地判别工程项目风险.但是,任何方法都有它的局限性,影响模型精度的因素可能来源于样本选取、码向量初值设定等几个方面.

参考文献(References)

[1] Straathof S M. Shannon's entropy as an index of product variety[J]. Economics Letters, 2007, 94(1): 297－303

[2] 李兴斯.一类不可微优化问题的有效解法[J].中国科学,A 辑,1994,24(4): 371－377
Li Xingsi. An efficiency solution of non-differentiable optimization problem [J]. Science in China, Ser A, 1994,24(4): 371－377(in Chinese)

[3] Janicak C A. Differences in relative risks for fatal occupational highway transportation accidents[J]. Journal of Safety Research, 2003, 34(5):539－545

[4] Ross K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems [J]. Proceedings of the IEEE, 1998, 86(11):2210－2239

(责任编辑:张　嵘)