Computer Science and Application 计算机科学与应用, 2011, 1, 63-68
http://dx.doi.org/10.12677/csa.2011.12013 Published Online September 2011 (http://www.hanspub.org/journal/csa/)

The Project Risk Estimate Model Based on the Projection Pursuit and Fuzzy Clustering

Shuhang Guo, Yan Li
School of Information, Central University of Finance and Economics, Beijing
Email: guoshuhang@hotmail.com; liyan36@126.com
Received: May 11th, 2011; revised: Jun. 27th, 2011; accepted: Jul. 5th, 2011.

Abstract: The objective of this paper is to construct a model which mainly based on projection pursuit model and use fuzzy estimates and clustering to raise the precision of the risk evaluation. Firstly, use projection pursuit model to compute the weight of each one of the risk factors, then check the deviation of each factor's score and use the interval point-rating method to reevaluate the score of these ones; secondly check the given scores of every expert and eliminate the deviated ones; finally, the risk of IT project is presented at the end for instance in steps, following that using projection pursuit model to verify the weight matrix to testify the feasibility of the model.

Keywords: Projection Pursuit Method; Fuzzy Estimate; Fuzzy Clustering; Project Risk Estimate

基于投影寻踪方法的模糊综合估计与聚类的工程项目风险评估

郭树行，李　妍
中央财经大学信息学院电子商务系，北京
Email: guoshuhang@hotmail.com; liyan36@126.com
收稿日期：2011 年 5 月 11 日；修回日期：2011 年 6 月 27 日；录用日期：2011 年 7 月 5 日

摘　要：本文基于投影寻踪的方法以模糊综合估计与聚类方法为核心，研究了工程项目风险量化评估问题。首先使用投影寻踪方法求出各个风险指标项的权重；接下来，检验各个指标项评分的偏离程度，对其中偏离度高的使用区间评分法；然后，检验各专家的评分，剔除其中偏离程度较大的专家评分项。最后通过 IT 项目实例计算该项目风险评分。

关键词：投影寻踪法；模糊估计；聚类分析；工程项目风险

1. 工程项目风险评估背景

在当前的时代，当众多的工程选择摆在企业面前时企业需要通过客观的风险评估方法去准确衡量并评价各个工程信息，以寻求可以使得企业得到最大运营效益的工程项目。而由于不同的项目工程所要评估的角度以及衡量标准，所以现今风险衡量已经成为了一个相对繁杂的过程。因此可以实现工程化的客观的风险评估显得尤为重要。

2. 工程项目风险研究综述

目前，依据项目所能提供的经验数据的多少及信息详细程度，风险评估技术可分为定性与定量两种。国内外对工程项目风险评估主要集中在以下几个方面：一是应用定性定量结合的方法进行风险评估，比如专家打分法、层次分析法和模糊数学法等；二是应用定量分析的方法进行风险评估，比如蒙特卡洛模拟、决策数算法基本原理与模型、贝叶斯网络；三是对 IT

项目风险的评估提前到项目投资决策阶段，应用期权理论来实现对项目风险评估；四是应用实证分析的方法，对影响 IT 项目成功的因素进行评估[1]。传统的分析方法都没有考虑风险因素之间的相互作用对风险评估的影响，比如采用层次分析和模糊数学方法时，是在假设风险之间是相互独立下进行的。

其中定量评价方法如蒙特卡洛模拟、决策数算法、贝叶斯网络等。蒙特卡洛模拟通过设定随机过程，反复生成时间序列，计算参数估计量和统计量究其分布特征。其技术的难点在于对风险因素相关性的辨识与评价，具有不确定性，该方法中所有的元素都同时受风险不确定性的影响。且系统的可靠性过于复杂，难以建立可靠简洁的数学模型；而决策数算法对记录大的数据库效果明显，可以在相对较小的计算量下处理变量/字段和决策树，清晰的显示哪些字段比较重要，免去很多数据预处理的工作。在发现市场关键驱动因素或者业务使用用户的关键特征方面非常有效；贝叶斯网络作为图论与概率论的结合，为变量间概率关系的图形化描述提供了一种将知识直观的图解可视化的方法，以贝叶斯概率理论为基础，具有成熟的概率推理算法和开发软件，为风险预测的贝叶斯模型建造和推理提供快捷的工具，加速了风险预测的有效性[1]。

3. 工程项目风险因素的确定

不同的工程项目需要从不同的角度进行评估，需要采用不同的风险评估指标。在确定工程项目风险因素时，可以通过不同工程项目所在领域所召开的年会上面所确定的风险评估单以及行业专家和企业管理人员的调查因素。

4. 基于模糊估计的 IT 项目风险评估机制

引入模糊聚类法的目的，就是在进行最终评估终值计算之前，对专家打分数据进行初始化处理，将某些与实际情况偏离较大的分数加以剔除，大大提高专家打分反映的真实程度，从而减少主观经验不同所造成的结果偏差[2]。

4.1. 传统聚类模糊分析及改进思路

聚类就是将物理或抽象对象的集合分为由类似的对象组成的多个类的过程，由聚类所生成的簇是一组数据对象的集合，这些对象与同一个类中的对象彼此相似，与其他类中的对象相异。传统的聚类分析是一种硬划分，它把每个待辨识的对象严格地划分到某类中，是一种“泾渭分明”的分类，但这种类别划分的界限是不合理的。在客观世界中，类与类之间往往存在着一个过渡性的边界，因此分类往往伴随着模糊性[3,4]。本文将模糊理论引入到聚类分析中，使分类显得更加合理，更符合实际客观情况，这就是模糊聚类分析。在信息系统安全评估中，不同类型的评估对象的安全属性各有不同，评价指标也各有侧重。传统的模糊聚类方法并没有考虑到评价指标之间的轻重关系，而是将其同等考虑，这直接影响了最终评估结果的准确性。改进的思路是在模糊综合分析过程中，引入评价指标的权重分析，通过 AHP(analytic hierarchy process)法计算评价指标之间权重的相对大小，并同时保证权重系数计算的客观性和准确性[2,5]。

4.2. 模糊聚类分析过程

4.2.1. 专家评分

分析专家评分数据,建立初始打分数据矩阵。论域 $U=\{x_1,x_2,\cdots,x_n\}$ 为被分类的对象，其中每个对象由 m 个数据指标表征，这样，建立起原始数据矩阵 $U=\left[x_{ij}\right]_{n\times m}$。

4.2.2. 确定权重向量

为减少在权重确定时的主观因素的影响本文选择使用投影寻踪法确定各个指标元素的权重向量。

此处投影寻踪法的总体思路是，将高维的数据投影到低维，用低维空间中的散点分布揭示高维数据的特征[1]。

这里，将通过构造非线性规划问题来获得能够代表高维数据的最佳投影方向[6]。

在模型的构造中，综合选取层次分析模型中的 10 个评价指标作为投影寻踪的方向指标，则

$$\boldsymbol{a}=\left(a_1a_2\cdots a_j\cdots a_{10}\right)\quad j=1,2\cdots,10$$

且 $\boldsymbol{a}$ 为单位长度向量，$a_j\in[0,1]$。将 X_{ij} 用上述方法无量纲化后得到的 Y_{ij} 投影到向量 $\boldsymbol{a}$ 上，所得的投影值即为关于 i 的投影指标函数[7]：

$$E(i)=\sum_{j=1}^{10}a_jY_{ij} \tag{1}$$

为了能让在低维下投影的散点能更好的代表高位数据的特征，在综合投影时，要尽可能多的获取 X_{ij} 的变异信息，要求投影的散点更为分散，即投影值的方差 S 要尽可能大[8]。

由此，可以构造出投影目标函数：$Q(a)=S$

其中

$$S=\sum_{i=1}^{n}\left(E(i)-\frac{\sum_{i=1}^{n}E(i)}{n}\right)^2 \tag{2}$$

为寻找最佳投影方向，就应该包含最多的 X_{ij} 的变异信息，也就是最大化投影值的标准差，因此构建有约束的非线性规划问题：

$$\max Q(a)=S$$

$$s.t.\quad \sum_{j=1}^{9}a_j^2=1\quad 0\le a_j\le 1 \tag{3}$$

4.2.3. 建立模糊关系矩阵

模糊关系的建立。模糊相似矩阵衡量分类数据之间的亲近程度。其中，$r_{pq}\in[0,1](p,q=1,2,\cdots,n)$ 表示分类对象 x_p 与 x_q 间的相似程度，r_{pq} 越小表示样本差异性越大，r_{pq} 越大表示样本差异性越小[4]。同时，由于相似系数 $r_{pq}=r_{qp}$，且对任意 p 都有 r_{pp}=1，即矩阵：

$$\boldsymbol{R}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}$$

满足对称性和自反性，主对角元素都为 1[9]。

计算 r_{pq} 值的确定方法大致分为3种：相似系数法、距离法以及贴近度法。相似系数法包括数量积法、相关系数法、夹角余弦法和指数相似系数法。距离法包括欧氏距离法、绝对值倒数法以及切比雪夫距离法等[7]。贴近度法包括算术平均最小法、最大最小法以及几何平均值最小法等。本文采用夹角余弦法确定 r_{pq} 取值[9]。将各样本作两两比较，每个样本的变量看作 k 维空间向量，然后计算彼此向量间夹角的余弦，计算公式为：

$$r_{pq}=\frac{\sum_{k=1}^{n}r_{pk}r_{qk}}{\sqrt{\sum_{k=1}^{n}r_{pk}^2}\sqrt{\sum_{k=1}^{n}r_{qk}^2}} \tag{4}$$

其中：k 表示每个样本有 k 个变量；r_{pk} 表示前一个样本在第 p 个变量上的取值；r_{qk} 表示后一个样本在第 q 个变量上的取值。

4.2.4. 建立模糊等价矩阵

由于模糊相似矩阵一般只满足自反性和对称性，并不满足传递性，所以需将模糊相似矩阵求解为模糊等价矩阵。本文采用传递闭包法，通过模糊数学的复合运算，实现多维数据聚类分析所需的对称性、自反性和传递性。通过求模糊相似矩阵的传递闭包，可以造一个模糊等价矩阵，即采用平方

$$\left(\boldsymbol{R}\times\boldsymbol{R}=\boldsymbol{R}^2\right):\boldsymbol{R}\to\boldsymbol{R}^2\to R^4\to\cdots\to R^{2k}\to\cdots\to\boldsymbol{t}(\boldsymbol{R})$$

在不超过 n 次运算后，当第 1 次出现 $R^{2k}=R^{2k+2}$ 时，R^{2k} 就是所求的传递闭包 $t(R)$[8]。

4.2.5. 模糊聚类计算

算出传递闭包后，选定不同的截取值 λ 对其进行截割分类。即对任意的 $\lambda\in[0,1],\left[\boldsymbol{t}(\boldsymbol{R})\right]^{\lambda}=\left(r_{pq}^{\lambda}\right)$ 为 $\boldsymbol{t}(\boldsymbol{R})$ 的 λ 截矩阵，其中：

$$r_{pq}^{\lambda}=\begin{cases}1 & r_{pq}>\lambda \\ 0 & r_{pq}<\lambda\end{cases}\qquad p,q=1,2,\cdots,n \tag{5}$$

当 $r_{pq}^{\lambda}=0$ 时，表示节点 p、q 不归为一类；当 $r_{pq}^{\lambda}=1$ 时，表示节点 p、q 归为一类。可以根据实际情况，选取不同的 λ 值，以便进行动态的聚类。

设专家 h 对风险控制效果状态的评价区间为 $[v_1,v_2]$，则 m 个专家的群体评价值取为($\bar{v}$)，其具体数值确定方法如下：$\bar{v}$ 是由专家打分后经过一定处理得出的。因为当指标由专家评判给出时，在很多情况下，专家很难给出一个确定的评价值，尤其是在指标的含义具有较大的模糊性时，专家更容易给出一个评价区间。为了使 $\bar{v}$ 的确定更具客观性，对专家给出的区间值做如下处理。

设有 k 个专家，第 h 个专家的评价区间为$[u_1, u_2]$，其中 $u_1\neq u_2$，若 $u_1=u_2=u^*$，则将 u^*按公式(1)(2)区间化处理成{ u_1^*， u_2^* }的形式。

$$u_1^*=u^*-\frac{1}{2k'}\sum_{n=1}^{k'}\left(u_2^{(h)}+u_1^{(h)}\right) \tag{6}$$

$$u_2^*=u^*+\frac{1}{2k'}\sum_{n=1}^{k'}\left(u_2^{(h)}-u_1^{(h)}\right) \tag{7}$$

式中 u_1^*，u_2^*分别表示当专家打分为单一值，即 $u_1(h)=u_2(h)=u^*$时，对 u^*区间化处理后的两个端点值；k'为专家打分为区间值(即 $u_1(h)\neq u_2(h)$)时的区间个数。

此时，对于 $u_1(h) = u_2(h) = u^*$来说，区间化后所选定的区间为$\{u_1^*, u_2^*\}$。根据集值统计方法，专家对某个指标的群体评价值取为：

$$\bar{u} = \frac{1}{2}\frac{\sum_{h=1}^{k}\left[\left(u_2^{(h)}\right)^2 - \left(u_1^{(h)}\right)^2\right]}{\sum_{h=1}^{k}\left[u_2^{(h)} - u_1^{(h)}\right]} \tag{8}$$

5. IT 项目实例

本文中选择工程项目中的 IT 项目实例作为分析对象。

首先专家对各个影响因素的重要程度进行评分，从中选择出十个重要程度较高的影响因素(在遇到评分相等的情况要尽量选择分属在不同的准则层下的指标变量)，并通过投影寻踪法对专家的评分进行处理，计算出各个风险指标项的权重。

再次实例中通过年会中制定的 IT 项目风险因素影响表确定初始风险影响因素，然后通过对该 IT 项目所在企业管理人员等的调研得出风险排名前十名的风险影响因素。继而通过投影寻踪法确定这十项风险影响指标的权重。在此次的调研结束后排名前十名的风险因素分别为：

1) 需求已经成为项目基准，但需求还在继续变化。

2) 风险管理粗心，导致未能发现重大的项目风险。

3) 在做需求文档中客户参与不够。

4) 管理层做出了打击项目组织积极性的决定。

5) 太不正规(缺乏遵循软件开发策略和标准的意识)，导致沟通不足，质量欠佳，甚至需重新开发。

6) 分别开发的模块无法有效集成，需要重新设计或制作。

7) 仅由管理层或市场人员进行技术决策，导致计划进度缓慢，计划时间延长。

8) 某些人员需要更多的时间适应还不熟悉的软件工具和环境。

9) 开发一种全新的模块将比预期花费更长的时间。

10) 客户的意见未被采纳，造成产品最终无法满足用户要求，因而必须重做。

专家对以上十个指标项进行评分(表 1)：

Table 1. Professor score
表 1. 专家评分

	指标 1	指标 2	指标 3	指标 4	指标 5	指标 6	指标 7	指标 8	指标 9	指标 10
专家 1	8	6	9	7	9	4	8	6	7	4
专家 2	5	3	3	4	9	7	9	3	4	3
专家 3	8	5	6	8	7	6	9	7	7	4
专家 4	8	5	8	6	8	5	7	5	5	5
专家 5	9	5	9	7	7	3	6	7	7	5
专家 6	7	7	5	5	8	5	8	7	6	6

由于模糊聚类分析法运算量较大，这里采用 SPSS 统计软件进行模糊聚类分析：首先对以上十个指标量各自的专家评分进行聚类分析，图 1~2 是使用 SPSS 进行聚类分析的冰柱图：

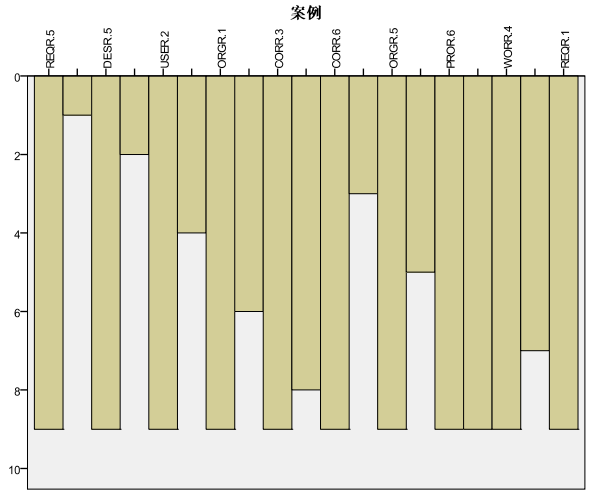

Figure 1. SPSS clustering icicle
图 1. SPSS 聚类分析冰柱图

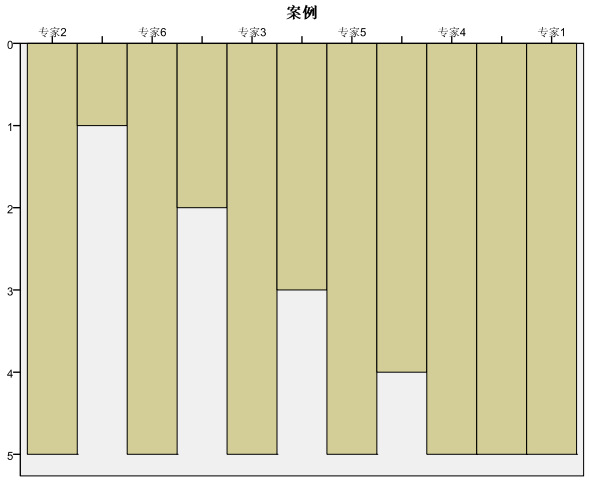

Figure 2. SPSS clustering icicle after adjustment
图 2. 调整后的 SPSS 聚类分析冰柱图

Table 2. REQR.5 adjusted professor score
表 2. REQR.5 区间评分法的调整值

指标 3	点 A	点 B	点 C
专家 1	9	6	7.5
专家 2	3	6	4.5
专家 3	5	7	6.0
专家 4	6	8	7.0
专家 5	7	9	8.0
专家 6	5	7	6.0

SPSS 聚类分析冰柱图中浅色的冰柱越长则代表该项的偏离程度越大，通过分析图 1 所示结果我们可以得出结论，指标 3 的偏离程度最大，表明专家在对此指标项的评分上面分歧较大，很难通过单一的确定数据确定该指标项的评分，则应使专家针对此指标项采取区间评分法。使用评分区间来对其偏离进行规避，对该项的评分调整如表 2。

接下来使用调整后的评分结果针对专家的评分进行聚类分析，结果如图 2 所示。

通过对图中数据的分析，可以发现专家 2 的评分与其他专家组成员相比偏离程度较大，即表明其评分较为主观，此处我们将专家 2 的评分去除掉，使得专家族的评分更加客观，更加具有参考性。

综上所述，上文中对数据的调整主要体现在两个方面：专家组对偏离程度较大的指标项进行区间评分法；剔除掉与整个专家评分组评分差异较大的专家评分项。

在对多个风险项目进行风险比较时可以选择使用投影寻踪法对数据进行处理，更加客观的设定判断矩阵，得出目标权重，实现对层次分析法的检验和补充。下面图 4 中显示出以上 10 个风险影响因素的权重，下面的 lingo 软件截图(图中 A(1)到 A(10)分别代表以上选取的 10 个指标项，顺序与以上十个指标项相同)中的权重基本与使用以上的层次分析法中计算出来的权重的趋势大致相同，如表 4 所示，以检验指标计算出的权重的客观性。

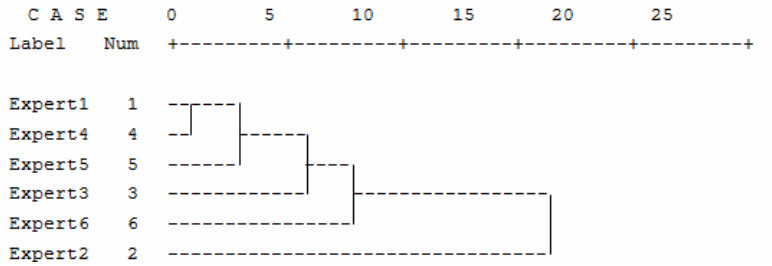

Figure 3. SPSS clustering after adjustment
图 3. 调整后的 SPSS 聚类分析树状图

Table 3. The adjusted professor score
表 3. 最终调整后的专家评分

	指标 1	指标 2	指标 3	指标 4	指标 5	指标 6	指标 7	指标 8	指标 9	指标 10
专家 1	5.7232	1.8612	2.817	1.953	2.5173	0.5368	1.7152	0.672	0.7889	0.3016
专家 3	5.7232	1.551	2.2536	2.232	1.9579	0.8052	1.9296	0.784	0.7889	0.3016
专家 4	5.7232	1.551	2.6292	1.674	2.2376	0.671	1.5008	0.56	0.5635	0.377
专家 5	6.4386	1.551	3.0048	1.953	1.9579	0.4026	1.2864	0.784	0.7889	0.377
专家 6	5.0078	2.1714	2.2536	1.395	2.2376	0.671	1.7152	0.784	0.6762	0.4524
风险得分						90.6089				

```
A( 1)       -0.7153850
A( 2)       -0.3101633
A( 3)       -0.3755679
A( 4)       -0.2789678
A( 5)       -0.2796545
A( 6)       -0.1342340
A( 7)       -0.2144482
A( 8)       -0.1120074
A( 9)       -0.1127268
A( 10)      -0.7538077E-01
```

Figure 4. Analysis result of pursuit projection model
图 4. Lingo 投影寻踪模型分析结果

Table 4. The weight table of the factors
表 4. 投影寻踪法计算出的指标权重表

	REQR.1	CORR.6	REQR.5	ORGR.5	CORR.3	DESR.5	ORGR.1	WORR.4	PROR.6	USER.2
投影寻踪法	0.7154	0.3102	0.3756	0.2790	0.2797	0.1342	0.2144	0.1120	0.1127	0.0754

6. 总结与展望

本文提出一种工程项目风险复合量化评估方法，此方法基于投影寻踪法以模糊综合估计与聚类方法为核心。通过企业 IT 项目实例计算该项目风险评分。实际结果表明该方法更加科学的剖析的 IT 项目风险状况。本文研究成果，为企业工程项目风险管理领域提供了进一步科学参考。

参考文献 (References)

[1] D. D. Chen, B. P. Ren. Analysis of China’s transitional economic performance: 1992-2006. Finance & Economics, CNKI:SUN: CJKX.0. 2009-05-012: 35-47.

[2] M. Filippone, F. CAmastra, F. Masulli, et al. A survey of kernel and spectrum methods for clustering. Pattern Recognition, 2008, 41(1): 176-190.

[3] D.-S. Chen, M.-C. Chen, and L.-L. Zhang. The research and application of clustering based on interval value. Mathematics in Practice and Theory, CNKI: SUN: SSJS.0. 2010-03-021: 193-235

[4] G.-L. Zhao, S.-R. Huang. Fuzzy clustering algorithm with modified kernel functions. Journal of Computer Applications, CNKI: SUN:JSJY.0.2010-07-065: 53-76

[5] I. Saha, U. Maulik. Fuzzy improved fuzzy clustering techniques for categorical data. AIP Conference Proceedings, 2009, 1089: 82-93.

[6] R. J. G. B. Campello, E. R. Hruschka, and V. S. Alvesvs. On the efficiency of evolutionary fuzzy clustering. Journal of Heuristics, 2009, 15(1): 43-76.

[7] C. G. Looney. Fuzzy connectivity clustering with radial basis kernel functions. Fuzzy Sets and Systems, 2009, 160(13): 1868 -1885.

[8] M.-S. Wu, X.-Z. Wu. Projection pursuit clustering method based on genetic algorithm. Statistics & Information Forum, CNKI: SUN:TJLT.0.2008-03-005: 107-119

[9] M. Lee, W. Pediycz. The fuzzy C-means algorithm with fuzzy P-mode prototypes for clustering objections having mixing features. Fuzzy Sets and Systems, 2009, 24(16): 3590-3600.