Journal of Risk Analysis and Crisis Response, Vol. 3, No. 2 (August 2013), 103-109

Decision-Making Model in the Environment of Complex Structure Data

Fusheng Yu
School of Mathematical Sciences, Beijing Normal University
Beijing, 100875, China
E-mail: yufusheng@bnu.edu.cn

Shihu Liu
School of Mathematical Sciences, Beijing Normal University
Beijing, 100875, China
E-mail: liush02@126.com

Received 14 June 2013

Accepted 28 June 2013

Abstract

For decision makers, the data property has a direct influence on the selection of decision making approaches and the reliability of decision results. Because of the complexity and diversity of practical decision data, some traditional decision approaches are not very good at reflecting the actual problem. For this, we propose a decision making model in the environment of complex structure data. The aim of this model is to discover the underlying community structure of the data by taking all aspects of original information into account. In this paper, the considered data is diversity, not only in structure but also in representation. What is more, a missing data compensation method is proposed by considering the information losing situation in practical decision making problem. Research shows that this model has great maneuverability. Especially, the proposed decision model seems more consistent with the actual decision problem, than decision model with single data structure.

***Keywords*:** Graph data, weighting determination, information fusion, completion of incomplete data, decision analysis.

复杂结构数据环境下的决策模型

于福生　刘士虎
北京师范大学/数学科学学院，北京 100875

摘要：对决策者而言，数据的特性直接影响决策工具的取舍和决策质量的可靠性。由于现实决策数据的复杂多变性，使得传统的一些决策模型不能够很好地实现决策目的。鉴于此，本文给出了一种基于复杂结构数据的决策模型。该模型旨在充分利用现有数据信息的基础上，实现挖掘数据潜在社团结构的目的。本文考虑的数据具有两大特性：一是结构的多样性，二是表示的多样性。甚至，考虑到实际问题中信息有损情况的存在，本文给出了一种有损数据补偿的方法。不难发现，该决策模型可操作性强，相对于单一数据格式下的决策模型，更贴近实际需要。

关键词：图数据，权重确定，信息融合，不完备数据完备化，决策分析

1. 引言

大到宇宙自然界，小到生活中的点点滴滴，风险无处不在，无时不有。这其中，有的风险是不以人的意志为转移的，如地震、台风、海啸等；有的风险是伴随着人类的活动而出现的，如工程建设风险、人员伤害风险、社会声誉风险等 [1, 2]。对于有人类参与的风险，风险的产生既有技术层面上的主观原因，又有社会层面上的客观原因。不管风险源自何处，归于何类，风险的最终承担者都是我们人类。于是，如何降低风险对人类的危害程度，是每一个风险管理者需要解决的首要问题。

为了最大程度地降低风险带来的损失，就需要对风险做出一个科学合理的评估。通常用定性或者定量的分析方法 [3] 研究风险发生的可能性及其后果的严重程度。但是常言道，天灾人祸不单行。对于实际问题的风险分析，往往存在诸多不确定性因素。诸如问题的描述、数据的采集、分析方法或工具的取舍，决策结果可靠性分析等。尤其对于问题的描述，一贯采用多指标描述的策略。但是，指标之间的交互影响，为问题的分析带来了困难。

简言之，不管问题有多么复杂，在风险分析过程中，通常是把抽象的问题模型化。对于不同的实际问题，可以结合具体的背景知识，建立相应的数学模型。然后利用该模型，对潜在的风险做一量化分析，在这个过程中对数据的处理显得至关重要。因为人们所面对的数据，往往具有复杂多变且不确定等诸多特点。甚者，对同一个问题，不同的数据分析员所面对的是不同的数据库，而且这些数据库之间是互相保密的。同时，不同的数据库，描述数据的方式也可能不相同。如有的数据库采用多指标描述法，有的采用关系描述法等。究其一点，数据类型大体上可以分为两类：向量型数据和关系型数据。通常，两种类型的数据是混合出现的。在风险分析的过程中，如何充分利用这两种不同类型的数据，对分析结果的可靠性有着巨大的影响。

基于此，本文对于具有复杂结构的数据的风险分析问题，从三个方面展开了相关研究。(1) 对于向量型数据中指标的权重确定方法的研究。熟知，对于多指标问题，不同的指标对于知识的认知程度，贡献通常是不一样的。故在分析过程中对于指标的区别看待是很有必要的。故在本文中我们给出了基于粒度的指标权重确定方法。同时，基于不同的模式对指标权重的贡献也不一样这一假设，我们给出了基于聚合算子的指标权重确定方法。(2) 对于数据损失的问题，给出了一种损失数据的补偿方法。在数据的采集、传输与存储过程中，出现数据的丢失是不可避免的现象。在此，基于数据的相似性，我们建立了一种基于“局部-整体”相似的丢失数据补偿方法。并用虚拟的知识表达系统，分析了该方法的可行性。(3) 对于复杂结构数据环境下的风险问题，我们从聚类的角度，做了相关分析。旨在挖掘出问题中潜在的“社团结构”或者“块结构”，为决策者的进一步决策提供一个指导作用。

作为本节的结束，接下来我们给出本文的基本框架。第 2 节简单的回顾一下本文需要的一些基本概念，如复杂结构数据、信息粒度、信息聚合与聚类分析等。第 3 节主要是对于向量型数据中指标权重的确定，从信息粒度和聚合算子两个方面给出确定方法。第 4 节是介绍向量型数据完备化的方法。第 5 节是利用聚类的思想，挖掘对于具有复杂结构数据的风险分析问题。

2. 预备知识

在该部分，我们对本文中用到的一些基本概念及其相关知识做一简要介绍。如：图数据，知识粒度，数据信息集成算子和聚类分析。对于其详细的介绍，可以参阅相关文献 [4,5]。

定义 1. (复杂结构数据)一个图数据可以表示为二元组 $G=(V,E)$，其中 V 代表该数据的模式集，E 代表模式之间的关系集。

对于一个只有有限个模式组成的图数据 G 而言，我们用 $V(G)=\{x_1,x_2,\cdots,x_n\}$ 表示顶点集，如果每个模式 x_i 是用 m 个指标来刻画的，则

$$V(G)=\begin{array}{c} \\ x_1 \\ \vdots \\ x_n \end{array}\begin{array}{c} \begin{array}{ccc} a_1 & \cdots & a_m \end{array} \\ \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix} \end{array}. \tag{1}$$

进一步，用 $E(G)$ 表示顶点关系集，即对于任意的两个对象 $x_i,x_j\in V(G)$，$(x_i,x_j)\in E(G)$ 当且仅当 x_i 和 x_j 存在某种关系。

显然，对于一个具有复杂结构的数据而言，对于不同的指标，其值域表示也不尽相同。同时，描述顶点之间关系的数据集 $E(G)$ 也未必只有一个。

定义 2. (知识粒度)给定向量型数据 $V(G)$，及在任意指标 a_i 下的划分 $P_i=\{p_{i1},p_{i2},\cdots,p_{in}\}$，则关于指标 a_i 的知识粒度定义为

$$G(a_i)=\frac{1}{n^2}\sum_{j=1}^{n}|\,p_{ij}\,|. \tag{2}$$

针对决策数据表示的异同性，许多学者提出了各种不同的数据集成算子 [5]，其一般的形式可以定义如下。

定义 3. (数据集成算子)设 $f:R^n\to R$，则称 $f_\omega(a_1,a_2,...,a_n)$ 为 $(a_1,a_2,...,a_n)\in R^n$ 的加权集成，其中 ω 为权系数且满足 $\omega_1+\omega_2+\cdots+\omega_n=1$.

显然，集成算子 f 以及权系数 ω 的取舍，依赖于决策者的决策需要以及具体的决策问题。

数据的聚类分析 [4]，旨在把给定的数据集 X 分成所期望的 k 类 $C=\{c_1,c_2,...,c_k\}$，使得类间尽可能远离，类内尽可能抱团。纵观现有的聚类算法，大致可以划分为硬划分(hard partition）和软划分(soft partition)两种。在硬划分下，每个模式只能属于某个特定的类。而在软划分下，对象是以一定的隶属度(介于 0 和 1 之间的某个数)归属于某个类。如对于向量型数据 X，划分成 k 类，最优聚类使下述误差函数

$$J(C,M)=\sum_{i=1}^{|X|}\sum_{c=1}^{k}\| x_i - v_c \|^2 \qquad (3)$$

达到最小值，其中 $M=\{v_1,v_2,...,v_k\}$ 为类的中心。相应的，其软划分，如经典的 FCM 算法，为求下述目标函数

$$J(C,M)=\sum_{i=1}^{|X|}\sum_{c=1}^{k}u_{ci}^{m}\| x_i - v_c \|^2 \qquad (4)$$

的最小值，其中 m 为模糊化因子。

在不引起歧义的前提下，下面符号在接下来的节中是通用的：G 代表图数据，E 代表图数据中顶点之间的关系，X 代表向量型数据，n 代表 G 或者 X 中顶点的个数，m 代表向量型数据的维数。在叙述的过程中，我们对向量型数据 X 和向量型数据 $V(G)$ 在符号表示上不做进一步的区别。

3. 决策模型中指标权重的确定

该部分我们主要讨论决策问题中，衡量决策目标的指标在决策过程中所扮演的角色。如果所有的指标所扮演的角色相同，则认为其对决策结果的贡献一样。这样，从权重的角度而言，所有指标的权重应该相等。一般而言，不同的指标对于决策的结果，所起的作用往往是不同的。于是，对不同的指标赋予不同的权重，对决策结果认知度的提高，就显得很有必要。决策的过程，实质上就是一个不断挖掘已有数据集中潜在信息的过程。下面，我们给出两种不同的权重确定方法，一是基于粒度思想的指标权重确定方法；二是基于聚合算子的指标权重确定方法。

3.1 基于粒度的权重确定方法

粒度这一概念，在特定层面上反映的是人类对知识的认知程度。对于一个向量型的数据，在某一指标下的粒度，可以刻画关于该指标对知识的认知能力。而且，如果粒度值越小，则代表该指标对知识的认知能力越强，反之，则认知能力越弱。

定义 4. 给定向量型数据集 X 以及关于任意指标 a_i 的粒度值 $G(a_i)$，则该指标 a_i 的权重定义为

$$\omega_i=\frac{1-G(a_i)}{\sum_{j=1}^{m}(1-G(a_j))}\ . \qquad (5)$$

显然，上式定义的指标的权重是归一化的。但是，对于某个特定的决策问题，往往考虑的指标很多。在此情况下，则有可能会出现很多指标的权重很小的现象。对此，我们给出一种基于权重的指标取舍方法。

定义 5. 给定向量型数据集 X 及任意指标 a_i 的权重 ω_i。对于事先假定的阈值 β，若存在指标 a_i 满足条件 $\omega_i<\beta$，则称该指标 a_i 是 β- 可删除的。

记 $\omega=(\omega_1,\omega_2,...,\omega_m)$ 是向量型数据集 X 中指标 $\{a_i\}_{i=1}^{m}$ 的权重，指标 a_i 是 β- 可删除的，以及 $\omega^*=\left(\omega_1^*,\omega_2^*,\ldots,\omega_{i-1}^*,\omega_{i+1}^*,\ldots,\omega_m^*\right)$，则根据定义 2 和定义 4，下面的结论显然成立。

命题 1. 对于指标 a_j $(i\neq j)$，成立 $\omega_j^*\geq\omega_j$。

证明：根据公式 (2) 可知，

$$\begin{aligned}\Delta\omega_j &= \omega_j^*-\omega_j \\ &= \frac{1-G(a_j)}{\sum_{l\neq i}(1-G(a_l))}-\frac{1-G(a_j)}{\sum_{l}(1-G(a_l))}\ , \\ &\geq 0\end{aligned} \qquad (6)$$

即 $\omega_j^*\geq\omega_j$。 □

命题 2. 对于指标 a_{j_1} 及 a_{j_2} $(i\neq j_1\neq j_2)$，若 $G(a_{j_1})<G(a_{j_2})$ 成立，则 $\omega_{j_1}^*-\omega_{j_1}\geq\omega_{j_2}^*-\omega_{j_2}$。

证明：由公式 (2) 及命题 1 可直接得证。 □

由上述命题可知：对于在阈值 β 下不可删除的某个指标，删除权重小于阈值 β 的指标后，它的权重不小于初始的权重值。还有，β- 不可删除指标关于知识认知能力的单调性是不变的。

例 1. 选取一些病人的看病记录，具体数据见下表 1. 用定义 4 给出的粒度的方法确定相应指标的权重。其中指标记为：头疼(a_1)、肌肉痛(a_2)、体温(a_3)、咳嗽(a_4)、睡眠质量(a_5)。

表 1. 一个关于某些病人的知识表达系统

病人	头疼	肌肉痛	体温	咳嗽	睡眠质量
1	否	是	正常	否	正常
2	是	是	正常	否	一般
3	是	否	偏低	否	正常
4	否	否	高	是	正常
5	是	是	很高	否	差
6	否	否	正常	是	正常
7	是	是	很高	是	差
8	否	是	较高	是	一般
9	否	是	高	否	一般
10	是	是	较高	否	差

我们约定：对于任意的指标 a_i，若 $x_{si}=x_{ti}$，则 x_{si} 和 x_{ti} 应该划分为一类。经计算可知关于指标 a_1 的粒度划分为 $P_1=\{\{1,4,6,8,9\},\{2,3,5,7,10\}\}$。根据定义 2 可知 $G(a_1)=0.5$。类似地，通过计算可知：

$$G(a_2)=0.22 \qquad G(a_3)=0.22$$
$$G(a_4)=0.52 \qquad G(a_5)=0.34$$

于是，由公式 (5) 可知，指标集$\{a_1,a_2,...,a_5\}$的权重向量为

$$\omega=(\omega_1,\omega_2,\omega_3,\omega_4,\omega_5)$$
$$=(0.1563,\ 0.2437,\ 0.2437,\ 0.1500,\ 0.2062)^{\cdot}$$

3.2 基于聚合算子的权重确定方法

在该部分，针对决策信息是不确定的情况，给出一种考虑决策目标偏好关系的指标权重确定方法。从 (1) 可以看出，决策信息x_{ij}一般都是事先通过一定的方法采集到的。实际问题的不确定性往往会导致数据x_{ij}表示的多样性与不确定性。在此，假设数据x_{ij}的不确定性由两部分构成：对目标x_i关于指标a_j的认可程度和否定程度。这样x_{ij}就可以表述为一个直觉模糊数[3] $x_{ij}=\left(\mu_{ij},\nu_{ij}\right)$。在不改变向量型数据集$X$符号表述的基础上，下面我们给出确定指标权重的相关定义与方法。

定义 6. 给定数据集X，则指标a_i基于聚合算子的权重定义为

$$\omega_i=\frac{agg(a_i)}{\sum_{j=1}^{m}agg(a_i)}\ . \qquad (7)$$

其中，$agg(a_i)$是数据$\{x_i\}_{i=1}^n$关于指标a_i的聚合值。

根据定义 3 可知，如果决策者看重整体数据的影响，则采用公式

$$agg\left(a_i\right)=1-\prod_{j=1}^{n}(1-\mu_{ji})^{\theta_j}+\prod_{j=1}^{n}\nu_{ji}^{\theta_j}\ . \qquad (8)$$

计算聚合值。若决策者想突出单个数据的作用，则采用公式

$$agg\left(a_i\right)=1+\prod_{j=1}^{n}\mu_{ji}^{\theta_j}-\prod_{j=1}^{n}(1-\nu_{ji})^{\theta_j}\ . \qquad (9)$$

计算聚合值，其中$\theta=(\theta_1,\theta_2,...,\theta_n)$为决策者对$\{x_i\}_{i=1}^n$的偏好，且满足归一化条件$\sum_{i=1}^{n}\theta_i=1$.

在文章5中，我们就对基于直觉模糊集的聚合算了做了一个简单的应用。当然，还存在诸多聚合算子，具体的表述可以参考文献 6和7.

例 2. 表 2 是一个关于 4 个病人的医疗诊断数据表。

下面我们计算关于 5 个体表特征的权重。在此取$\theta=(0.25,0.25,0.25,0.25)$，即对 4 个病人同等看待。

经计算，基于公式 (8)的特征权重为

$\omega_1=0.2121$　　$\omega_2=0.2136$

$\omega_3=0.1872$　　$\omega_4=0.2046$

$\omega_5=0.1825$

从上述两种确定权重的方法不难发现，基于粒度的权重确定方法，侧重于从分类的角度，考察关于某个指标对知识认知能力的强弱。而基于聚合算子的权重确定方法，考虑了决策者对决策目标的偏好程度。总之，两种方法各有优点。但是同时存在一个缺点，就是如果决策问题的指标个数很多，则指标之间的差异就有可能变得很小。对于此类问题，我们在接下来的部分，将从数据变换的角度，展开详细研究

4. 决策模型中有损数据的完备化

虽然发达的科技能够为数据的采集与存储提供便捷的方式，但是对于实际问题，在数据采集、传输与存储的过程中，出现部分信息损坏或丢失的情况是不可避免的。接下来我们从一个全新的角度，对缺失数据给出一种补偿的方法。该方法未必是最好的，但从下面的实例可以看出，它是行之有效的。

问题描述：给定向量型数据X，存在部分信息丢失的情况，假定所有指标的权重是已知的。

有损信息补偿方法：

1. 把存在有损信息的对象归为一类：

$$MS=\left\{x_i\,\middle|\,x_{ij}=*,\ j\in\{1,2,\cdots m\}\right\}. \qquad (10)$$

2. 计算对象x_i和x_j关于指标c_p的局部相似性。数据为实数时：(1) x_{ip}和x_{jp}均存在，若 $x_{ip}=x_{jp}$，则$ls(i,j,p)=1$，否则$ls(i,j,p)=0$. (2) 对于其余的任何情况，取$ls(i,j,p)=0.5$. 数据为集值时：(1) 若x_{ip}和x_{jp}均存在，则

$$ls(i,j,p)=\frac{|x_{ip}\bigcap x_{jp}|}{|x_{ip}\bigcup x_{jp}|};\qquad (11)$$

否则$ls(i,j,p)=1$. (2)其余取$ls(i,j,p)=0.5$.

3. 计算任意两个对象之间的整体相似度

表 2. 医疗诊断问题

病人	Temperature	Headache	Stomach pain	Cough	Chest pain
Al	(0.8,0.1)	(0.6, 0.1)	(0.2,0.8)	(0.6,0.1)	(0.1,0.6)
Bob	(0.0,0.8)	(0.4, 0.4)	(0.6,0.1)	(0.1,0.7)	(0.1,0.8)
Joe	(0.8,0.1)	(0.8, 0.1)	(0.0,0.6)	(0.2,0.7)	(0.0,0.5)
Ted	(0.6,0.1)	(0.5, 0.4)	(0.3,0.4)	(0.7,0.2)	(0.3,0.4)

$$gs(i,j)=\sum_{p=1}^{m}\alpha_p\times ls(i,j,p)\ . \tag{12}$$

4. 对于存在信息丢失的任意对象 x_i，计算

$$C_i=\left\{x_j\,\middle|\,(x_i,x_j)\in T\right\}, \tag{13}$$

其中 $T=\left\{\left(x_i,x_j\right)\middle|gs(i,j)\ge\xi\right\}$。

5. 对于 $x_i\in MS$，如果 x_i 缺失的信息是关于 c_p 的，先计算 $C_p(x_i)=\left\{x_{tp}\,\middle|\,x_t\in C_i,x_{tp}\ne *\right\}$，然后(1)若 c_p 是实数型的，则 $C_p(x_i)=[\wedge C_p(x_i),\vee C_p(x_i)]$；(2)若 c_p 是集值型的，则 $C_p(x_i)=\bigcup x_{tp}$。

例 3. 表 3 是一个包含 5 个对象和 6 个指标的一个不完备信息系统。

聚类分析的基本思想，已经被广泛应用到风险分析中，诸如震后灾情评估[11]，财务风险分析[12]，风险投资[13] 等诸多领域。在此我们就实际问题的复杂性，从聚类的角度作进一步分析研究。

对于一个实际问题，除了可以利用多个指标来刻画一个对象的特性外，所要研究的对象之间往往还存在千丝万缕的关系。这种关系，构成了所谓的关系数据。于是，对于实际问题的建模，就是对于一个具有复杂结构数据 $G=(V,E)$ 的再分析过程。其中，V 代表所要研究的对象集，是一向量型数据；E 代表该数据对象之间的关系集，是一关系型数据。综上所述，对于一个实际问题的分析，就转变

表 3. 一个不完备的信息系统

编号	c_1	c_2	c_3	c_4	c_5	c_6
1	3	2	*	*	{1, 3}	{3}
2	2	*	[0.0, 1.5]	[0.5, 1.0]	{1, 2}	*
3	1	2	*	[1.5, 2.0]	{1, 2, 3}	*
4	*	0	[1.0, 2.0]	*	{1, 2}	{2, 3}
5	2	1	[0.5, 1.5]	[0.3, 1.8]	{3}	{1, 2}

表 4.完备化的不完备信息系统

编号	c_1	c_2	c_3	c_4	c_5	c_6
1	3	2	**[0.5, 1.5]**	**[0.3, 2.0]**	{1, 3}	{3}
2	2	**[0,2]**	[0.0, 1.5]	[0.5, 1.0]	{1, 2}	{1, 2, 3}
3	1	2	**[0.0, 2.0]**	[1.5, 2.0]	{1, 2, 3}	{2, 3}
4	**[1,2]**	0	[1.0, 2.0]	**[0.5, 2.0]**	{1, 2}	{2, 3}
5	2	1	[0.5, 1.5]	[0.3, 1.8]	{3}	{1, 2}

显然，第 3 节介绍的权重的确定办法是无法确定相应指标的权重的。在此不妨假定所有指标的权重是相等的，即 $\alpha_i=1/6$.通过一系列的计算，丢失数据补偿后的信息系统表示为 4. 从中可以看出，该方法是可行的。比如对指标 c_3 而言，关于 x_1 的补偿完全不等同于 x_3 的补偿值。这也说明，对于该方法，不再是单一的取最大值，最小值或者平均值来补偿丢失的数据。

5. 基于聚类的复杂结构环境下的决策模型

我们知道，风险分类[8-10] 是风险管理中的一项基本工作。它是根据不同标准，对已知风险进行分类，旨在提高风险管理效率的基础上降低风险管理成本。

成对多结构数据的分析。

通常情况下，所研究的问题都只是对 $V(G)$ 或者 $E(G)$ 的单独分析。分析过程简单明了，但是处理结果有失偏颇。为了避免这一点，最直接的方法就是对 $V(G)$ 和 $E(G)$ 中的数据实现对位加权处理。权重往往是通过以往实验或者经验值估计出来的。

对于任意两个对象 x_i，$x_j\in V(G)$，不妨记 d_{ij} 为关于向量型数据集 $V(G)$ 的不相似描述，e_{ij} 为关于关系型数据集 $E(G)$ 的不相似描述，下面我们给出一种不加权的信息融合方法。

定义 7. 给定向量型数据 $V(G)$ 和关系型数据 $E(G)$，则称

$$t_{ij}=\left(d_{ij}+1\right)\left(e_{ij}+1\right)-1 \tag{14}$$

为对象 x_i 和 x_j 无权的信息融合值，记为 T.

显然，上述信息融合矩阵 T 满足以下性质：

命题 3. 对于任意 x_i, $x_j \in V(G)$，成立 $t_{ij} = t_{ji}$。

命题 4. 对于任意 $x_i \in V(G)$，成立 $t_{ii} = 0$。

显然，融合后的数据 T 是一个关系型数据，描述的是任意两个对象之间的不相似程度。从本质上而言，和最初的关系型数据 $E(G)$ 在数据表示上没有本质的区别。可是在数据信息的蕴含方面，T 却包含了向量型数据 $V(G)$ 所描述的信息，要比 $E(G)$ 蕴含的信息丰富。

对于一个具体的风险分析问题，一旦 $V(G)$ 和 $E(G)$ 给定，就可以由公式 (14)，计算出融合后的关系型数据 T。接下来，我们就需要对关系型数据 T，挖掘其潜在的社团结构，对数据分析人员对问题的进一步研究提供有用的信息。

问题描述：对于某个风险分析问题，其模型化的数据表示为 $G = (V, E)$。希望把 $|V(G)|$ 个对象划分成 k 个团体，使得团体之间的差异尽可能的大，但是团体内部差异尽可能的小。

模型建立：

1. 向量型数据集 $V(G)$ 的处理：如果 $V(G)$ 存在部分信息丢失的情况，则利用第 4 节给出的有损信息补偿方法，补全丢失的信息。如果 $V(G)$ 不存在信息丢失的情况，则利用第 3 节介绍的方法，确定向量型数据 $V(G)$ 中相应指标的权重。
2. 利用定义 7 介绍的方法，把 $V(G)$ 和 $E(G)$ 融合成新的关系型数据 T。其中上一步计算出来的指标的权重，作用于关于向量型数据集 $V(G)$ 中任意对象的不相似性度量方面。
3. 根据问题的需要，利用谱聚类的方法，将关系型数据 T 划分成需要的 k 类。

由上可知，对于具体的风险决策问题，由采集到的数据，上述模型可以实现进一步挖掘该问题中潜在“社团”结构的目的。相对于利用单一结构数据寻找“社团”而言，复杂数据结构环境下的决策模型，提供的结果更可靠，更具有说服力。这不仅能为对该数据的进一步分析提供良好的指导，同时也能降低接下来处理数据时人为带入的不确定性。

例 4. 表 5 是一个复杂结构数据 $G = (U, V)$ 的向量型数据集 $V(G)$ 的表示，表 6 是对应的关系型数据 $E(G)$ 的表示。

表 5：向量型数据 $V(G)$

编号	c_1	c_2	c_3	c_4	c_5	c_6
1	14.2	1.7	2.4	11.4	127	2.8
2	13.2	1.7	2.1	14.0	100	2.6
3	13.1	2.3	2.6	18.6	101	2.8
4	14.3	1.9	2.5	16.8	113	3.8
5	13.2	2.5	2.8	21.0	118	2.8
6	14.2	1.7	2.4	15.2	112	3.2
7	14.2	1.8	2.4	14.6	96	2.5
8	14.0	2.1	2.6	17.6	127	2.6
9	14.0	1.6	2.1	14.0	96	2.8
10	13.8	1.3	2.2	16.0	98	2.9

表 6：关系型数据 $E(G)$

编号	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	0	0	1	1	0	0
2	1	0	1	1	0	1	0	0	0	0
3	0	1	0	0	1	0	1	1	0	0
4	0	1	0	0	0	0	1	1	0	0
5	0	0	1	0	0	0	0	0	0	0
6	0	1	0	0	0	0	1	1	0	0
7	1	0	1	1	0	1	0	0	0	0
8	1	0	1	1	0	1	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	0	0

在表 6 中， 数值 1 表示对象 x_i 和对象 x_j 恒不相似，即 $e_{ij} = 1$,反之， 0 表示对象 x_i 和 x_j 绝对相似，即 $e_{ij} = 0$.接下来我们从三个方面对模型加以说明：

（I）只考虑数据 $V(G)$

利用经典的 k-means 聚类算法把向量型数据集 $V(G)$ 划分成 $k = 3$ 类：

$$\begin{aligned} c_1 &= \{x_2, x_3, x_7, x_9, x_{10}\} \\ c_2 &= \{x_4, x_5, x_6\} \\ c_3 &= \{x_1, x_8\} \end{aligned} \quad . \tag{15}$$

（II）只考虑数据 $E(G)$

利用谱聚类算法(Spectral clustering)把关系型数据集 $E(G)$ 仍然划分成 $k = 3$ 类：

$$\begin{aligned} \mathrm{c}_1 &= \{x_2, x_7\} \\ \mathrm{c}_2 &= \{x_1, x_3, x_4, x_5, x_6, x_9, x_{10}\} \\ \mathrm{c}_3 &= \{x_8\} \end{aligned} \quad . \tag{16}$$

（III）同时考虑数据 $V(G)$ 和 $E(G)$

首先计算各指标的权重，在这里我们采用基于粒度的权重确定方法。经计算可知

$\omega = (0.1619, 0.17, 0.1619, 0.1781, 0.1741, 0.1538)$.

然后在计算关于 $V(G)$ 的加权不相似性度量。在此基础上，根据定义 7，计算数据集 $V(G)$ 和 $E(G)$ 无权融合后的信息 T。由于 T 在结构表示上仍然是一个关系型数据，故采用谱聚类算法 (Spectral clustering)，可知当 $k = 3$ 时，得到的划分为

$$c_1 = \{x_2, x_7, x_9, x_{10}\},$$
$$c_2 = \{x_1, x_3, x_4, x_5, x_6\} \quad (17)$$
$$c_3 = \{x_8\}$$

从式子 (15)-(17) 不难发现，对象 x_4, x_5 和 x_6 一致划分在同一类，对象 x_2 和 x_7 一致划分在同一类。相对于综合考虑两种结构的数据而言，在只考虑向量型数据集 $V(G)$ 时，原本划分在同一类的对象 x_1 和 x_3，将被划分到其余不同的类中。而当只考虑关系型数据 $E(G)$ 时，对象 x_9 和 x_{10} 将被划分到其余的同一类中。

6. 结论

风险无处不在，无时不有。在面对具有复杂结构数据的决策问题时，进一步挖掘其潜在的数据结构，能够为降低风险带来巨大的帮助。本文正是基于这一点，首先探讨了相应指标的权重确定方法。其次，对于数据丢失的情况，建立了一种数据补偿的方法，并用相应的实例验证了该方法的可行性。对于问题的数学模型所面对的是一个具有复杂结构的数据时，我们利用聚类的思想挖掘其潜在的社团结构，期望达到进一步降低问题处理中带来的不确定性。

参考文献

1. C. F. Huang, Natural Disaster Risk Analysis and Management. (Science Press, Beijing, 2012).
   黄崇福.自然灾害风险分析与管理[M].北京:科学出版社，2012.
2. L. B. Tao, Y. S. Li, Z. L. Feng, et al., Project risk analysis theory and practice.(Tongji University Press, Shanghai, 2006).
   陶履彬，李永盛，冯紫良等. 工程风险分析理论与实践[M]. 上海:同济大学出版社，2006.
3. C. H. Le, H. Y. Ding, G. H. Dong, et al., Risk analysis of failure damage to marine riser based on fuzzy fault tree, Journal of Natural disasters,21(2)(2012) 173-179.
   乐丛欢，丁红岩，董国海等.基于模糊故障树的海洋立管破坏失效风险分析[J]. 自然灾害学报，2012,21(2): 173-179.
4. W. Pedrycz, Knowledge-Based Clustering: from data to information granules. (John Wiley & Sons, New Jersey, 2005).
5. S. H. Liu, F. S. Yu, Aggregation operators based MCDM with intuitionistic fuzzy information, in Proceedings of the fifth annual meeting of risk analysis council of China association for disaster prevention, eds. C. F. Huang and G. F. Zhai (Atlantis Press, Paris, 2012), pp. 411-416.
6. Z. S. Xu, Intuitionistic fuzzy information aggregation theory and application. (Science Press, Beijing, 2008).
   徐泽水.直觉模糊信息集成理论及其应用[M].北京:科学出版社，2008.
7. Z. S. Xu, Intuitionistic fuzzy aggregation operators, IEEE Transactions on Fuzzy Systems, 15(6) (2007) 1179-1187.
8. P. J. Shi, C. F. Huang, T. Ye, et al., Constructing China's comprehensive risk management system, Disaster Reduction in China, 1(2)(2005), 164-167.
   史培军，黄崇福，叶涛等.建立中国综合风险管理体系[J].中国减灾, 2005,1(2): 35-37.
9. C. F. Huang, A trapezoid framework for integrated risk management, Journal of Natural Disasters, 14(6)(2005) 9-14
   黄崇福.综合风险管理的梯形架构[J].自然灾害学报，2005,14(6): 9-14.
10. P. J. Shi, T. Ye, J. A. Wang, et al., Integrated governance of natural disaster risk, Journal of Beijing Normal University,5(2006), 130-136.
    史培军，叶涛，王静爱等.论自然灾害风险的综合行政管理[J].北京师范大学学报，2006, 5: 130-136.
11. L. G. Tian, Y. Li, Fuzzy cluster analysis in the application of reservoir evaluation after earthquake, yellow river, 32(1)(2010),130-131.
    田林刚，李洋.模糊聚类分析在震后水库风险评价中的应用[J].人民黄河，2010, 32(1): 130-131.
12. X. G. Zhou, R. Zhu, Analysis of corporate financial risk based on fuzzy clustering and pattern recognition, Science and technology management research, 8(2012),115-123.
    周晓光，朱荣.基于模糊聚类和模式识别的企业财务风险分析[J].科技管理研究，2012, 8: 115-123.
13. Y. P, Yang, J. Wang, Research on industry clustering of venture capital in China, Science and technology management research, 12(2012), 164-167.
    杨艳萍，王静. 我国风险投资的行业聚类研究[J]，科技管理研究，2012, 12: 164-167.