

# 基于数据资源的认知图挖掘方法

陈 庄<sup>1)</sup> 阿里·蒙特瑟密<sup>2)</sup>

<sup>1)</sup>(重庆工学院计算机科学与工程学院 重庆 400050)

<sup>2)</sup>(麦克斯特大学商学院 加拿大安大略省汉密尔顿市 L8S 4M4)

**摘 要** 认知图(Cognitive Map, CM)是一种新型知识管理方法,具有直观的知识表达能力、强大的基于数字矩阵的推理机制等优点.但是,要充分展示CM的这些优点,首先必须获得正确的CM图.传统的获取CM图的方法(如问卷法、头脑风暴法、样本学习法等)常常借助专家经验,因其过分强调主观因素而忽视客观数据资源,容易导致信息丢失现象.为此,该文提出了一种基于客观数据资源来挖掘CM图的新方法,它主要由数据库初始化技术、权重系数优化方法、CM图的简化策略等组成.实验结果表明:该方法能挖掘出构成CM图的所有节点间的所有关系,并可针对这些关系间的重要程度对CM图作适当简化;与传统的方法相比,该方法所挖掘出的CM图具有更为丰富的信息.

**关键词** 认知图;数据挖掘;神经网络;数据库;学习算法

中图法分类号 TP34

## The Methodology of Mining Cognitive Maps Based on Data Resources

CHEN Zhuang<sup>1)</sup> Montazemi Ali R.<sup>2)</sup>

<sup>1)</sup>(School of Computer Science and Engineering, Chongqing Institute of Technology, Chongqing 400050)

<sup>2)</sup>(Business School, McMaster University, Hamilton ON L8S 4M4, Canada)

**Abstract** Cognitive Map (CM) is a new kind of method of knowledge management, which has many advantages such as: it is relative easy to use for representing structured knowledge, the inference mechanism can be computed by numeric matrix operation, etc. However, in order to exhibit these advantages about CM, the first step is that the corrected CMs must be obtained. Traditional approaches for obtaining the CMs, including questionnaire method, brainstorming method, and sample learning method, always rely on experience of domain experts. Because these methods put much emphasis on the subjective factors and neglect the objective data resources, they always lose some information. Therefore, this paper proposes a new methodology of mining the CMs based on data resources, which mainly includes database preprocessing technology, optimization algorithm for weight coefficients, and simplification strategy for CMs. The experimental results show that: the new method can mine all possible relationships among all nodes to form the CMs, and can also simplify it according to the significant degree of those relationships; the CMs mined by the new method has more information than the CMs obtained by traditional approaches.

**Keywords** cognitive map; data mining; neural network; database; learning arithmetic

## 1 引 言

认知图(Cognitive Map, CM)是一种知识管理

方法和知识表示工具<sup>[1,2]</sup>,是对相关领域中的实体之间联系的“因果表达”<sup>[3]</sup>.认知图的概念最早是由Tolman于1948年提出的,旨在描述一种心理学模型<sup>[4]</sup>.1976年, Axelord在其名著《Structure of De-

cision》中最先将认知图应用于决策领域,用 CM 图建立了有关社会、经济及政治领域中知识和决策模型<sup>[1]</sup>.1986 年,Kosko 等人在 Axelord 的基础上,将模糊关系引入到 CM 中,提出了模糊认知图(Fuzzy Cognitive Map,FCM)<sup>[5-6]</sup>.

认知图(包括模糊认知图)是模糊逻辑、神经网络、图论等学科的交叉和综合<sup>[2,7]</sup>.作为一种新的智能方法,认知图与传统的人工智能方法既有联系又有本质的区别:(1)从图的结构上,CM 可以视为单层的神经网络,因而神经网络的方法可以借鉴.但同时又比神经网络有着较强的语义,更易于建立直观的结构化知识模型,并对所建立的模型有着更强的可解释性<sup>[6,8-10]</sup>;(2)从图的特征上,CM 又可分为一种有向图,因此可借助图论方面的理论对其进行深入研究<sup>[11-13]</sup>;(3)从推理机制上,CM 可通过矩阵的运算来进行推理,可以充分利用矩阵方面的有关理论,比常规的“基于规则的推理”运算,其内涵更为丰富;(4)CM 允许反馈机制的存在,这也为复杂系统建模提供了可能<sup>[2,7]</sup>.正因为 CM 有着这些优点,因而越来越受到学术界的重视,并已经逐渐成为人工智能领域的新的发展方向.

但是,要展示 CM 的上述优点,其首要的工作和关键的任务就是要获取正确或准确的相关领域的 CM 图.现有的获取 CM 图的方法(如问卷法、头脑风暴法、样本学习法等)主要借助专家经验,因其过分强调主观因素而忽视客观数据资源,容易导致信息丢失现象.为此,本文将提出一种获取 CM 图的

新方法,即基于客观数据资源的 CM 图挖掘方法,并以某金融数据库为工程对象开展实验研究,验证此方法的有效性.

## 2 相关文献分析

CM 的表示方法主要有三种,即有向图法(directed graph)<sup>[1]</sup>、关联矩阵法(valency matrix)<sup>[1-2]</sup>和神经网络法(neural network)<sup>[14-15]</sup>.从本质上来看,这三种对认知图的描述方法是等价的.图 1(a)、(b)、(c)是针对某城市社会问题的 CM 图三种不同描述<sup>[1]</sup>.图中,节点  $N_1, N_2, N_3, N_4, N_5, N_6, N_7$  代表领域(domain)中的相关概念(concepts)或变量(variables),在本例中这些节点分别代表某城市社会问题:城市人口量、移民数量、现代化程度、垃圾量、卫生设施、疾病人口量以及细菌量(单位面积);弧(arcs)或权重数(weight)则代表节点之间的因果关系(causal relationships),它是一种三值(+1、-1、0)逻辑关系:“+1”表示节点间是“正”因果关系(即原因节点与结果节点的变化方向一致),“-1”表示节点间是“负”因果关系(即原因节点与结果节点的变化方向相反),“0”表示节点间无因果关系;这些三值(+1、-1、0)逻辑关系直接作为相关元素而构成关联矩阵.要充分发挥 CM 的优势,正确、准确地获取 CM 图(尤其是节点数及其间的因果关系)至关重要,因此 CM 图的获取技术一直受到许多学者的关注.目前,主要有三类获取 CM 图的方法,即问卷调查法(questionnaire method)<sup>[16]</sup>、头脑

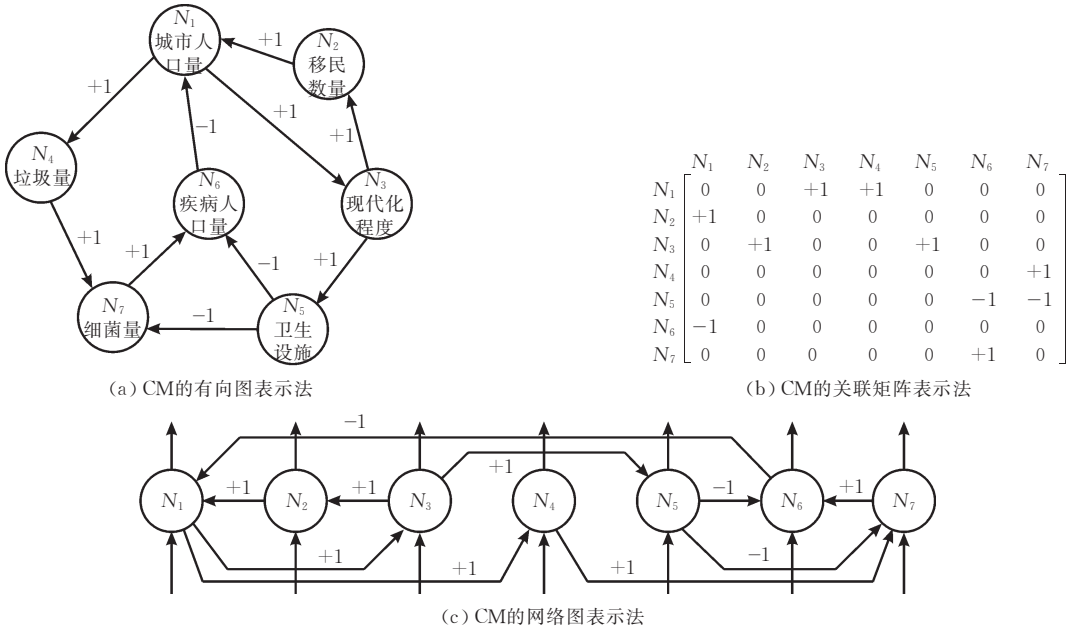


图 1 CM 的三种表示方法

风暴法 (brainstorming method)<sup>[17]</sup> 和样本学习法 (sample learning method)<sup>[18-21]</sup>. 其中, 问卷调查法<sup>[16]</sup>是通过问卷表方式向领域专家进行调查, 通过专家的反馈意见来确定节点的个数以及节点间的因果关系(弧值); 头脑风暴法的基本思想是<sup>[17]</sup>: 邀请 5~10 位专家参与讨论会, 讨论会的主题就是“确定节点的个数以及节点间的因果关系”. 在会上, 让所有与会专家自由发表自己的意见, 充分发挥其创造性思维, 互相启迪, 互相影响, 求大同、存小异, 最终构思出认知图; 样本学习法则形式较多, 但其基本思想是相似的, 即: 在通过专家确定出算法初始值和节点数的前提下, 选择少量典型的样本数据及相应的数学模型进行学习, 并由此确定出节点间的因果关系(弧值). 例如, Schnceider 提出一种比较学习方法, 即通过比较两组变量(节点)的相似性和变化的趋势, 来决定其弧值<sup>[18]</sup>; Huerga 提出了一种微分学习算法, 该算法是基于一种更新规则——某个概念(节点)的变化依赖于在同一时间内所有概念对它改变的影响<sup>[19]</sup>; Papageorgiou 等人提出一种非线性 Hebbian 学习算法, 该算法是利用 Hebbian 规则建立非线性优化模型, 并利用最速下降法进行学习<sup>[20]</sup>; Parsopoulos 等人提出粒子群学习算法, 该方法是通过时间序列数据的学习来求邻接矩阵<sup>[21]</sup>.

上述三类方法主要是依靠专家经验, 因而从本质上来讲是属于“主观性”方法. 这些主观方法主要存在两点不足: (1) 它们基本上不与领域的数据库产生联系或仅与少量的样本数据产生联系, 忽视了组织内(或行业领域)的大量数据资源; (2) 它们所依赖的专家本身存在知识局限性、观点的偏向性、决策的人为性等不足. 因此, 利用这些方法所形成的 CM 图存在丢失信息情况是在所难免的.

3 认知图挖掘方法

3.1 认知图挖掘方法的总体结构

3.1.1 基本思路

根据上述介绍的 CM 概念及其表示方法, 作下面的合理假设:

- (1) 客观数据资源, 经整理(即利用下文的数据库初始化技术)后, 共有  $m$  个节点  $N_1, N_2, \dots, N_m$ ;
- (2) 每个节点  $N_j (j=1, 2, \dots, m)$  均受其它节点  $N_i (i \neq j)$  的影响(即  $N_i$  与  $N_j$  产生“因果关系”), 其影响程度为  $w_{ij}$ . 若某节点  $N_{i_0}$  对  $N_j$  不产生影响, 则  $w_{i_0, j}=0$ .
- (3) 排除节点对自身的影响, 即认为节点与其

自身不产生因果关系, 此时  $w_{ii}=0 (i=1, 2, \dots, m)$ .

基于这些合理假设, 可以将  $m$  个节点的 CM 图视为类似“一层神经网络图”<sup>[7, 12]</sup>, 如图 2 所示. 图中,  $m$  个节点  $N_1, N_2, \dots, N_m$  可以视为  $m$  个神经元.

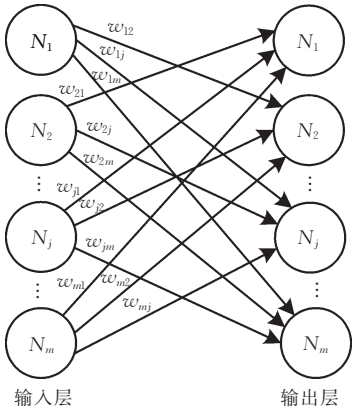


图 2 CM 的一层神经网络来描述

根据图 2, 易得到下述邻接矩阵 (adjacency matrix):

$$W = \begin{bmatrix} 0 & w_{12} & \cdots & w_{1, m-1} & w_{1, m} \\ w_{21} & 0 & \cdots & w_{2, m-1} & w_{2, m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{m-1, 1} & w_{m-1, 2} & \cdots & 0 & w_{m-1, m} \\ w_{m, 1} & w_{m, 2} & \cdots & w_{m, m-1} & 0 \end{bmatrix} \quad (1)$$

式中, 邻接矩阵  $W$  的元素  $w_{ij}$  又称为权重系数 (weight coefficient), 因而邻接矩阵也称作权重系数矩阵<sup>[22]</sup>.

邻接矩阵  $W=(w_{ij})$  与图 1 所示的关联矩阵  $V=(v_{ij})$  关系密切, 依据下式可直接将邻接矩阵转换成关联矩阵<sup>[23]</sup>

$$v_{ij} = \begin{cases} 1, & w_{ij} > 0 \\ 0, & w_{ij} = 0 \\ -1, & w_{ij} < 0 \end{cases} \quad \begin{matrix} (i = 1, 2, \dots, n; \\ j = 1, 2, \dots, n) \end{matrix} \quad (2)$$

这样, 获取认知图的问题便转换成获取邻接矩阵的问题, 即如何获取邻接矩阵  $W$  的  $n(n-1)$  个元素值的问题.

3.1.2 总体结构

基于上述基本思路, 给出获取 CM 图的总体结构, 如图 3 所示. 图中, “数据库初始化”的功能在于整理不规范、不标准甚至零乱的“客观数据资源”, 从中提炼出关键的属性变量(即字段变量, 图中为  $m$  个节点  $N_1, N_2, \dots, N_m$ )及其规范化值(即数据库记录值); “神经元激活函数”的功能是模拟人的认知过程, 它可以选择以下三种函数

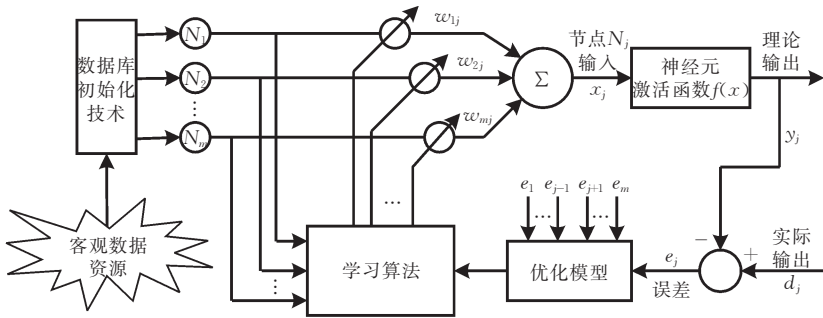


图 3 获取 CM 图的总体结构

$$f(x) = \frac{1}{1 + e^{-x}} \text{ (S 函数)} \tag{3a}$$

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \text{ (双曲正切函数)} \tag{3b}$$

$$f(x) = x \text{ (线性函数)} \tag{3c}$$

变量  $x_j, y_j$  分别为神经元  $N_j$  的输入、输出(又称理论输出),  $d_j$  为实际输出(即数据库的记录值),  $e_j$  为实际输出与理论输出间的误差, 该误差与其它神经元( $N_1, \dots, N_{j-1}, N_{j+1}, \dots, N_m$ )经认知后产生的误差( $e_1, \dots, e_{j-1}, e_{j+1}, \dots, e_m$ )一并构成优化模型的输入. 我们的目标就是通过“学习算法”来调整权重系数  $w_{ij} (i, j=1, 2, \dots, m, i \neq j)$ , 从而使“优化模型”的目标函数达到最小.

3.2 数据库初始化技术

数据库的初始化技术较多, 在此仅介绍三种常用的技术: 标准化法、归一法、小数缩放法. 在具体应用时, 可根据领域或组织的具体情况选择其中一种.

3.2.1 标准化法

标准化法是一种统计方法, 它是通过一定的非线性变换, 将原始的数据库记录值转换成“标准化数据”(standardization data). 标准化法是按下式对原始数据进行变换

$$d_j^{(i)} = \frac{o_j^{(i)} - u_j}{s_j} \quad (j = 1, 2, \dots, m; i = 1, 2, \dots, n) \tag{4a}$$

式中,  $o_j^{(i)}$  为原始数据库的第  $j$  个字段变量的第  $i$  条记录值;  $d_j^{(i)}$  为新的标准化数据库中的第  $j$  个字段变量的第  $i$  条记录值;  $m, n$  分别为原始(或新)数据库的字段变量个数和记录个数;  $u_j, s_j$  分别为原始数据库第  $j$  个字段变量的所有记录的平均值和标准差, 它们分别由下面两式得到

$$u_j = \frac{1}{n} \sum_{i=1}^n o_j^{(i)},$$
$$s_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (o_j^{(i)} - u_j)^2}.$$

3.2.2 归一法

归一法是通过一定的线性变换进行转换, 将原始的数据库记录值转换成  $[0, 1]$  之间的数据. 其转换算式为

$$d_j^{(i)} = \frac{o_j^{(i)} - L_j}{M_j - L_j} \quad (j = 1, 2, \dots, m; i = 1, 2, \dots, n) \tag{4b}$$

式中,  $M_j, L_j$  分别为原始数据库第  $j$  个字段变量的所有记录中的最大值和最小值, 它们分别由下面两式得到

$$M_j = \max \{ o_j^{(1)}, o_j^{(2)}, \dots, o_j^{(n)} \},$$
$$L_j = \min \{ o_j^{(1)}, o_j^{(2)}, \dots, o_j^{(n)} \}.$$

3.2.3 小数缩放法

小数缩放法是原始的数据库记录值的小数点进行前或后移动, 以保证其绝对值之最大值介于  $[0.1, 1)$  之间.

设  $A_j$  为原始数据库第  $j$  个字段变量的所有记录中的绝对值最大值, 即

$$A_j = \max \{ |o_j^{(1)}|, |o_j^{(2)}|, \dots, |o_j^{(n)}| \}$$

则小数缩放法的转换算式为

$$d_j^{(i)} = o_j^{(i)} \times 10^r \quad (j = 1, 2, \dots, m; i = 1, 2, \dots, n) \tag{4c}$$

式中,  $r$  为满足下式的整数值(含负整数、零或正整数)

$$0.1 \leq A_j \times 10^r < 1.$$

3.3 权重系数优化方法

利用图 3 及上述数据库初始化技术所得到的规范化数据, 以下给出权重系数优化方法, 包括优化模型和学习算法.

3.3.1 优化模型

优化模型的基本思想是: 使得所有节点(即  $N_1, N_2, \dots, N_m$ )经神经元激活函数所产生的理论输出与实际输出(即标准数据库字段变量的记录值)之间的误差达到最小. 基于该思想, 得到下述优化模型

$$\min g(\mathbf{W}) = \sum_{i=1}^m \sum_{j=1}^n (d_j^{(i)} - y_j^{(i)})^2 \tag{5}$$

式中,  $\mathbf{W}$  为式(1)的邻接矩阵, 其元素(即权重系数)为待优化的变量;  $m, n, d_j^{(i)}$  的定义同式(3);  $x_j^{(i)}, y_j^{(i)}$  分别为节点  $N_j$  神经元激活函数的输入及其输出, 其计算公式如下

$$x_j^{(i)} = \sum_{t=1}^{j-1} w_{tj} d_t^{(i)} + \sum_{t=j+1}^m w_{tj} d_t^{(i)} \quad (6)$$

$$y_j^{(i)} = f(x_j^{(i)}) \quad (7)$$

3.3.2 学习方法

学习方法即优化方法, 旨在通过相关迭代算法找到最优的权重系数  $w_{ij}$ , 使得算式(5)的目标函数值最小. 以下给出学习方法的基本步骤.

1. 初始化:
  - 1.1. 从标准数据库中读取所有字段变量的记录值  $d_j^{(i)}$  ( $j=1, 2, \dots, m; i=1, 2, \dots, n$ );
  - 1.2. 从式(3a)、(3b)、(3c)中选择一种神经元激活函数, 以下统一用式(3);
  - 1.3. 设定学习系数  $\rho$ , 其值为较小的正实数;
  - 1.4. 设定权重系数的初始迭代值  $w_{ij}^{(0)}$  ( $i, j=1, 2, \dots, m; i \neq j$ ).
  - 1.5. 设定迭代循环变量  $q$  的初始值,  $q=0$ .
  2. 循环计算下述子步 2.1、2.2、2.3、2.4、2.5、2.6, 直至满足相关条件跳出循环:
    - 2.1. 根据式(6)计算  $x_j^{(i)}$  ( $j=1, 2, \dots, m; i=1, 2, \dots, n$ );
    - 2.2. 根据式(3)、(7)计算  $y_j^{(i)}$  ( $j=1, 2, \dots, m; i=1, 2, \dots, n$ );
    - 2.3. 依据下式计算梯度值  $\delta_{ij}$ :
$$\delta_{ij} = -\frac{\partial g(\mathbf{W})}{\partial w_{ij}^{(q)}} \quad (8)$$
    - 2.4. 依据下式更新权重系数  $w_{ij}^{(q+1)}$ :
$$w_{ij}^{(q+1)} = w_{ij}^{(q)} + \rho \delta_{ij} \quad (9)$$
    - 2.5. 根据式(5)计算  $g(\mathbf{W})$  值;
    - 2.6. 循环判断: 若  $g(\mathbf{W})$  (或  $\delta_{ij}$ ) 足够小, 则跳出循环, 转步 3; 否则,  $q=q+1$ , 转步 2.1.
  3. 输出最终的权重系数  $w_{ij}^{(q)}$  ( $i, j=1, 2, \dots, m; i \neq j$ ), 即得到权重系数的优化值. 至此, 优化迭代结束.

3.4 CM 图的简化策略

利用上述的初始化方法、优化方法及式(2), 便可以得到 CM 图, 它包括了所有节点间的所有关系. 尽管这些节点及其间的关系对于 CM 的研究十分重要, 但是在实际应用中, 往往不需要这么复杂的 CM 图. 因此, 应遵循一定的策略对上述获得的 CM 图进行简化.

为此, 笔者根据文献[22-23]的“设置阈值生成 CM 图”的方法, 来简化 CM 图. 具体简化思想是: 根据“阈值”来判断节点间的“关系”的“影响程度”, 若其“影响程度”超过“阈值”, 则认为该关系重要, 应在

CM 图中保留; 否则, 说明的该“关系”不重要, 应在 CM 图中省略之.

基于上述简化思想, 首先对于节点  $N_i (i=1, 2, \dots, m)$ , 根据实际情况设置与其相对应的“阈值”  $T_i^n, T_i^p (i=1, 2, \dots, m)$ ; 然后, 针对所获得的邻接矩阵权重系数  $w_{ij} (i, j=1, 2, \dots, m; i \neq j)$  (这些权重系数可以视为节点间的关系“影响程度”), 按“阈值”来判断其重要性. 这样, 依据下式便可以将邻接矩阵  $\mathbf{W}$  转换成简化后的关系矩阵  $\mathbf{V}$ :

$$v_{ij} = \begin{cases} 1, & w_{ij} > T_i^p \\ 0, & -T_i^n \leq w_{ij} \leq T_i^p \\ -1, & w_{ij} < -T_i^n \end{cases} \quad (i=1, 2, \dots, n; j=1, 2, \dots, n) \quad (10)$$

显然, 就同一个邻接矩阵  $\mathbf{W}$  而言, 按式(10)所得到的 CM 图比按式(2)所得到的 CM 图要简单一些.

4 实验案例

以下以某大型数据库——固定收入贸易数据库 (Fixed-Income Trades Database, FITD) 为例, 说明如何利用上述算法从中挖掘 CM 图.

4.1 问题描述

FIDB 数据库是路透社 (Reuters) 为加拿大多家金融、证券公司 (包括 Toronto Stock Exchange, Bank of Montreal Nesbitt Burns Inc., CIBC World Market Inc., RBC Dominion Securities Inc. 等) 提供金融服务的在线、实时大型数据库. 它包括了近年来这几家金融企业的固定收入的相关贸易数据, 现已有 200 多个字段属性变量和百万余条记录.

基于 FIDB 数据资源, 为了分析影响“有价证券”(bond) 的相关“因素”及其间的“因果关系”, 我们抽取了 FIDB 中的 40 种有价证券在 2004.3.11~2006.8.7 期间的相关数据, 形成 FIDB 数据库的一个子库 SFIDB. 在数据库 SFIDB 中, 共有 22785 条记录, 且每个记录均包括了下述 6 个属性变量:

- (1) Price: 证券价格. 为了简化操作, 用变量  $C_1$  表示(下同).
- (2) Yield: 证券到期的年收益率, 用变量  $C_2$  表示.
- (3) Credit quality: 发行证券商的诚信度, 用变量  $C_3$  表示.
- (4) Issue size: 证券类别, 用变量  $C_4$  表示.
- (5) Coupon rate: 证券半年红利率, 用变量  $C_5$  表示.

(6) Time to maturity: 证券到期期限,用变量  $C_6$  表示.

至此,该实验的目的演变为:从包括 22785 记录、6 个变量的数据库 SFIDB 中挖掘 CM 图,从而为相关金融机构对“有价证券”进行分析、决策提供依据.

表 1 数据库经初始化后的结果(仅列举其中一小部分)

记录号	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$
1	0.7682	1.6607	1.6607	1.1063	-0.0151	1.33912
2	0.1437	-2.1157	-2.1157	-0.0846	0.6647	-1.6026
3	-0.0503	-0.5863	-0.5863	-0.0846	0.6647	-0.8671
4	-0.1889	-0.8601	-0.8601	-0.0846	-2.5215	-0.1317
5	1.0804	-1.5587	-1.5587	-0.0846	0.6647	-0.1317
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
22781	-0.1962	0.3862	-1.4918	-1.2755	0.2859	0.6037
22782	-1.0584	1.8401	0.1065	1.1063	-2.2354	0.6037
22783	-1.4144	0.6505	0.2425	-0.0846	0.2859	-0.8671
22784	-1.0628	0.3956	-1.4335	-1.2755	-0.0151	-1.6026
22785	-1.2481	0.4994	-0.3167	-1.2755	1.3438	-0.8671

4.2.2 实验方法及实验结果

笔者利用 Matlab 7.1 编程实现了 3.3 节的优化方法,并基于表 1 的数据得到了认识图的邻接矩阵  $W$ :

$$W = \begin{bmatrix} 0 & -0.836 & -0.322 & 0.545 & 0.765 & 0.671 \\ -0.648 & 0 & -0.254 & 0.254 & 0.584 & 0.757 \\ -0.098 & -0.099 & 0 & 0.040 & 0.180 & 0.017 \\ 0.160 & 0.096 & 0.038 & 0 & -0.164 & -0.126 \\ 0.894 & 0.880 & 0.693 & -0.654 & 0 & -0.457 \\ 0.267 & 0.388 & 0.022 & -0.171 & -0.156 & 0 \end{bmatrix}$$

(11)

4.2 实验过程及结果分析

4.2.1 初始化

经过反复实验并依据金融专家经验对其结果进行比较后,我们选择了标准化方法,对 SFIDB 数据库中的所有数据进行初始化.表 1 列举了利用式(3)对 SFIDB 进行初始化后的部分结果.

根据式(11)及(2),得到关联矩阵  $V$ :

$$V = \begin{bmatrix} 0 & -1 & -1 & +1 & +1 & +1 \\ -1 & 0 & -1 & +1 & +1 & +1 \\ -1 & -1 & 0 & +1 & +1 & +1 \\ +1 & +1 & +1 & 0 & -1 & -1 \\ +1 & +1 & +1 & -1 & 0 & -1 \\ +1 & +1 & +1 & -1 & -1 & 0 \end{bmatrix}$$

(12)

根据式(12)的关联矩阵,便得到 CM 图,如图 4 所示.

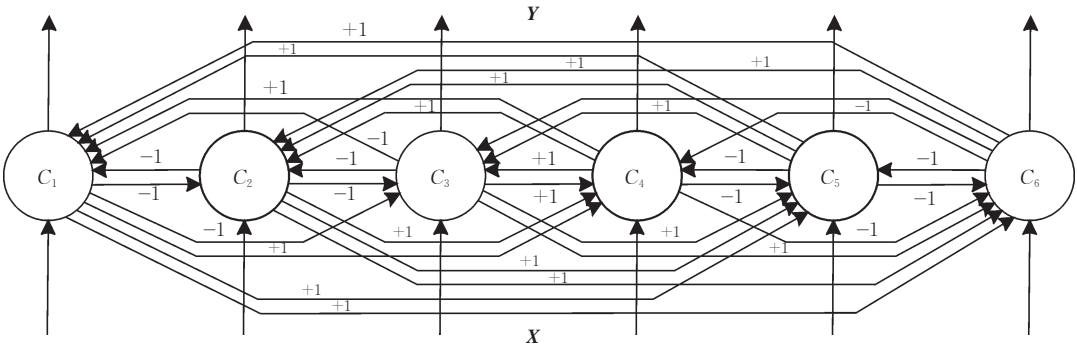


图 4 基于数据库 SFIDB 的 CM 图

4.2.3 实验结果分析

针对 4.1 节描述的问题,依据专家经验所得到的 CM 图如图 5 所示<sup>[24-26]</sup>.显然,基于数据库 SFIDB 挖掘出来的 CM 图(图 4)包含了所有节点间的所有关系,而图 5 仅包含部分节点间的因果关系,所以图 4 所包含的信息当然比图 5 的信息丰富.

尽管如此,在实际应用中,往往不需要这么复杂的 CM 图,即不需要所有节点间的因果关系,而是

仅保留重要的因果关系.以下,笔者利用基于专家经验的图 5 的特征和 3.4 节提出的简化策略对图 4 进行简化,并比较简化后的 CM 图与图 5 间的区别.

图 4 的邻接矩阵——式(11),其元素  $a_{ij}$  ( $i, j = 1, 2, \dots, 6; i \neq j$ ) 说明了节点  $C_i$  ( $i = 1, 2, \dots, 6$ ) 对其它节点  $C_j$  ( $j = 1, 2, \dots, i-1, i+1, \dots, 6$ ) 的影响程度,即因果关系  $C_i \rightarrow C_j$  的重要程度.



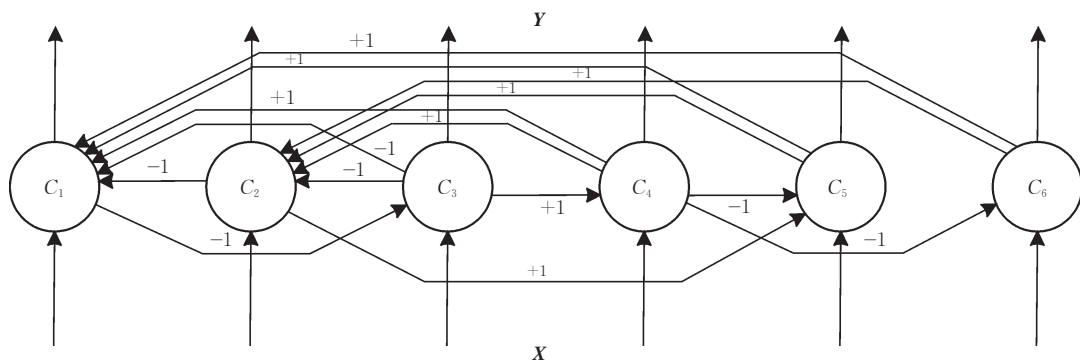


图 5 基于专家经验的 CM 图

(1) 对该矩阵的第 1 行而言, 由于专家 CM 图仅保留了一个负因果关系  $C_1 \rightarrow C_3$ , 而没有任何正因果关系. 因此, 为保留负因果关系  $C_1 \rightarrow C_3$ , 应设置“阈值”  $T_1^n$ , 其值属于区间  $[0, |-0.322|)$  即可; 为去除所有正因果关系, 应设置“阈值”  $T_1^p$ , 其值属于区间  $[0.765, +\infty)$  即可. 这样, 将第 1 行的“阈值”设为  $T_1^n = 0, T_1^p = 0.765$ .

(2) 对该矩阵的第 2 行而言, 由于专家 CM 图保留了一个负因果关系  $C_2 \rightarrow C_1$  和一个正因果关系  $C_2 \rightarrow C_5$ , 因此, 为保留负因果关系  $C_2 \rightarrow C_1$ , 应设置“阈值”  $T_2^n$ , 其值属于区间  $[|-0.25|, |-0.648|)$  即可; 为保留正因果关系  $C_2 \rightarrow C_5$ , 应设置“阈值”  $T_2^p$ , 其值属于区间  $[0.254, 0.584)$  即可. 这样, 将第 2 行的“阈值”设为  $T_2^n = 0.254, T_2^p = 0.254$ .

(3) 对该矩阵的第 3 行而言, 由于专家 CM 图保留了仅有的两个负因果关系  $C_3 \rightarrow C_1, C_3 \rightarrow C_2$  和一个正因果关系  $C_3 \rightarrow C_4$ , 因此, 为保留负因果关系  $C_3 \rightarrow C_1, C_3 \rightarrow C_2$ , 应设置“阈值”  $T_3^n$ , 其值属于区间  $[0, |-0.098|)$  即可; 为保留正因果关系  $C_3 \rightarrow C_4$ , 应设置“阈值”  $T_3^p$ , 其值属于区间  $[0.017, 0.040)$  即可. 这样, 将第 3 行的“阈值”设为  $T_3^n = 0, T_3^p = 0.017$ .

(4) 对该矩阵的第 4 行而言, 由于专家 CM 图保留了仅有的两个负因果关系  $C_4 \rightarrow C_5, C_4 \rightarrow C_6$  和两个正因果关系  $C_4 \rightarrow C_1, C_4 \rightarrow C_2$ , 因此, 为保留  $C_4 \rightarrow C_5, C_4 \rightarrow C_6$ , 应设置“阈值”  $T_4^n$ , 其值属于区间  $[0, |-0.126|)$  即可; 为保留正因果关系  $C_4 \rightarrow C_1, C_4 \rightarrow C_2$ , 应设置“阈值”  $T_4^p$ , 其值属于区间  $[0.038, 0.096)$  即可. 将第 4 行的“阈值”设为  $T_4^n = 0, T_4^p = 0.038$ .

(5) 对该矩阵的第 5 行而言, 由于专家 CM 图没有任何负因果关系和保留了两个正因果关系  $C_5 \rightarrow C_1, C_5 \rightarrow C_2$ , 因此, 去除所有负因果关系, 应设置“阈值”  $T_5^n$ , 其值属于区间  $[|-0.654|, |-\infty|)$  即可; 为保留正因果关系  $C_5 \rightarrow C_1, C_5 \rightarrow C_2$ , 应设置“阈值”

$T_5^p$ , 其值属于区间  $[0.693, 0.880)$  即可. 这样, 将第 5 行的“阈值”设为  $T_5^n = 0.654, T_5^p = 0.693$ .

(6) 对该矩阵的第 6 行而言, 由于专家 CM 图没有任何负因果关系和保留了两个正因果关系  $C_6 \rightarrow C_1, C_6 \rightarrow C_2$ , 因此, 去除所有负因果关系, 应设置“阈值”  $T_6^n$ , 其值属于区间  $[|-0.171|, |-\infty|)$  即可; 为保留正因果关系  $C_6 \rightarrow C_1, C_6 \rightarrow C_2$ , 应设置“阈值”  $T_6^p$ , 其值属于区间  $[0.022, 0.267)$  即可. 这样, 将第 6 行的“阈值”设为  $T_6^n = 0.171, T_6^p = 0.022$ .

依据上述所设置的“阈值”  $T_i^n, T_i^p (i=1, 2, \dots, 6)$  及算式(10), 得到简化后的关联矩阵  $V$  为

$$V = \begin{bmatrix} 0 & -1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & +1 & +1 \\ -1 & -1 & 0 & +1 & +1 & 0 \\ +1 & +1 & 0 & 0 & -1 & -1 \\ +1 & +1 & 0 & 0 & 0 & 0 \\ +1 & +1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (13)$$

由式(13)易得到简化后的 CM 图, 如图 6 所示. 比较图 6 与图 5, 不难看出: 图 6 比图 5 多了三个“关系”(见图 6 中虚线关系), 而这三个关系正是专家 CM 图中丢失的重要信息. 以下简要分析之.

(1) 新增关系  $C_1 \rightarrow C_2$  的分析. 在专家图 5 中, 保留了关系  $C_1 \rightarrow C_3$ , 其权重系数为  $-0.322$ ; 而关系  $C_1 \rightarrow C_2$  的权重系数为  $-0.836$ , 且  $|-0.836| > |-0.322|$ , 说明  $C_1$  对  $C_3$  的影响比  $C_1$  对  $C_2$  的影响大, 即关系  $C_1 \rightarrow C_3$  比关系  $C_1 \rightarrow C_2$  重要. 因此, 既然保留了关系  $C_1 \rightarrow C_2$ , 那么比之更重要的关系  $C_1 \rightarrow C_3$  理应保留.

(2) 新增关系  $C_2 \rightarrow C_6$  的分析. 在专家图 5 中, 保留了关系  $C_2 \rightarrow C_5$ , 其权重系数为  $0.584$ ; 而关系  $C_2 \rightarrow C_6$  的权重系数为  $0.757$ , 且  $0.757 > 0.584$ , 说明关系  $C_2 \rightarrow C_6$  比关系  $C_2 \rightarrow C_5$  重要. 因此, 既然保留了关系  $C_2 \rightarrow C_5$ , 那么比之更重要的关系  $C_2 \rightarrow C_6$  理应保留.

(3) 新增关系  $C_3 \rightarrow C_5$  的分析. 在专家图 5 中, 保留了关系  $C_3 \rightarrow C_4$ , 其权重系数为 0.040; 而关系  $C_3 \rightarrow C_5$  的权重系数为 0.180, 且  $0.180 > 0.040$ , 说

明关系  $C_3 \rightarrow C_5$  比关系  $C_3 \rightarrow C_4$  重要. 因此, 既然保留了关系  $C_3 \rightarrow C_4$ , 那么比之更重要的关系  $C_3 \rightarrow C_5$  理应保留.

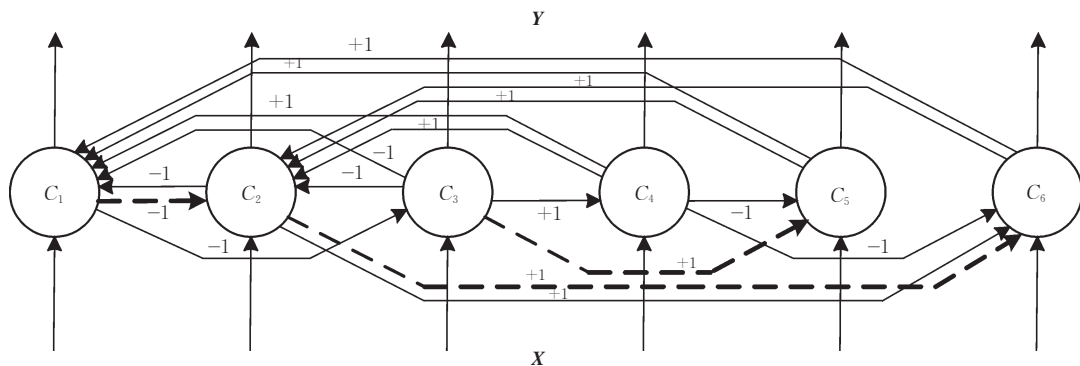


图 6 基于数据库 SFIDB 的简化 CM 图

以上分析表明, 基于数据资源挖掘出的 CM 图与依据专家经验所得到的 CM 图相比, 不仅所包含的信息更丰富, 而且还能协助专家找回其所丢失的重要信息.

为了建立和完善基于客观数据资源的挖掘 CM 图的方法体系, 下一步我们将进一步开展 CM 推理机制研究以及 CM 图的评价体系研究.

## 参 考 文 献

## 5 结 语

本文针对传统的获取 CM 图的方法存在的不足, 提出一种新的基于客观数据资源的挖掘 CM 图的方法, 并开展了实验研究, 具体贡献如下:

(1) 分析了 CM 的图的研究背景及相关获取认知图的方法的特点及不足, 并由此提出了基于数据资源的认知图挖掘方法的总体结构, 它主要由客观数据资源、数据库初始化技术、神经元激活函数、优化模型及学习算法构成;

(2) 探讨了数据库初始化技术, 包括标准化法、归一法、小数缩放法, 并给出了这些技术的数学模型描述;

(3) 建立了权重系数优化方法, 包括优化模型及学习算法. 其中, 优化模型充分利用了数据库中的每一个客观数据; 而学习算法则旨在搜索一组权重系数, 从而使认知结果与实际结果的偏差最小.

(4) 提出了 CM 图的简化策略, 该策略主要针对通过优化方法所获得权重系数, 按“阈值”来判断其重要性, 并在 CM 图中仅保留重要的因果关系;

(5) 利用 Matlab 7.1 实现了上述的技术、方法和策略, 并以某大型金融数据库 FIDB 为工程背景, 进行了实验研究, 验证了本文方法的有效性, 说明了利用本方法所获得的 CM 图与基于专家经验得到的 CM 图相比, 不仅所包含的信息更丰富, 而且还能协助专家找回其所丢失的重要信息.

- [1] Axelrod R. Structure of Decision. Princeton: Princeton University Press, 1976
- [2] Eden C. Cognitive mapping: A review. European Journal of Operational Research, 1988, 36(1): 1-13
- [3] Montazemi A R, Conrath D W. The use of cognitive mapping for information requirements analysis. MIS Quarterly, 1986, 10(1): 44-55
- [4] Tolman E C. Cognitive maps in rats and man. Psychological Review, 1948, 55(4): 189-208
- [5] Kosko B. Fuzzy cognitive maps. International Journal of Man-Machine Studies, 1986, 24(1): 65-75
- [6] Kosko B. Neural Networks and Fuzzy Systems. New Jersey: Prentice-Hall, 1992
- [7] Aguilar J. A survey about fuzzy cognitive maps papers. International Journal of Computational Cognition, 2005, 3(2): 27-33
- [8] Caudill M. Using neural nets: Fuzzy cognitive maps. AI Expert, 1990, 5(6): 49-53
- [9] Aguilar J. A dynamic fuzzy-cognitive-map approach based on random neural networks. International Journal of Computational Cognition, 2003, 1(4): 91-107
- [10] Tsadiras A K, Margaritis K G. Cognitive mapping and certainty neuron fuzzy cognitive maps. Information Sciences, 1997, 101(1): 109-130
- [11] Miao Y, Liu Z Q, Siew C K, Miao C Y. Dynamical cognitive network—an extension of fuzzy cognitive map. IEEE Transactions on Fuzzy Systems, 2001, 9(5): 760-770
- [12] Zhang W R. Equilibrium relations and bipolar cognitive mapping for online analytical processing with applications in international relations and strategic decision support. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 2003, 33(2): 295-307



- [13] Zhang W R. NPN fuzzy sets and NPN qualitative algebra: A computational framework for bipolar cognitive modeling and multiagent analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 1996, 26: 561-574
- [14] Anderson J A. Neural models with cognitive implications//Laberge B, Samuels S J. *Proceedings of the Basic Processes in Reading Perception and Comprehension*. New Jersey, 1977: 27-90
- [15] Lippmann R P. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 1987(4): 12-35
- [16] Roberts F S. The questionnaire method//Axelrod R. *Proceedings of the Structure of Decision*, Princeton: Princeton University Press, 1976: 333-342
- [17] Hong T, Han L. Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks. *Expert Systems With Applications*, 2002, 23(1): 1-8
- [18] Schncider M, Shnaider E, Kandel A. Constructing fuzzy cognitive maps//*Proceedings of the 1995 IEEE International Conference on Fuzzy Systems*. Yokohama, Japan, 1995: 2282-2288
- [19] Huerga A V. A balanced differential learning algorithm in fuzzy cognitive maps. *Universitat Politècnica de Catalunya (UPC)*, Spain: Technical Report, 2002
- [20] Papageorgiou E, Stylios C, Groumpos P. Fuzzy cognitive map learning based on nonlinear Hebbian rule//*Proceedings of the Australian Conference on Artificial Intelligence*. Australian, 2003: 256-268
- [21] Parsopoulos K E, Papageorgiou E I, Groumpos P P, Vrahatis M N. Fuzzy cognitive maps learning using particle swarm optimization. *Intelligent Information Systems Archive*, 2005, 25 (1): 95-121
- [22] Wang S. A dynamic perspective of differences between cognitive maps. *The Journal of the Operational Research Society*, 1996, 47(40): 538-549
- [23] Smith K L, Wirth A. Measuring differences between cognitive maps. *The Journal of the Operational Research Society*, 1992, 43(12): 1135-1150
- [24] Huang C, Montazemi A R. Review about cognitive maps and its application in data mining. *McMaster University, Canada: Research Report*, 2007
- [25] Fabozzi F J. *Bond markets, analysis and strategies* (5th edition). New Jersey: Pearson Education, Inc., 2004
- [26] Khurana I K, Raman K K. Are fundamentals priced in the bond market. *Contemporary Accounting Research*, 2003, 20 (3): 465-494
- [27] Montazemi A R, Gupta K M. On the effectiveness of cognitive feedback from an interface agent. *OMEGA*, 1997, 25 (6): 643-658
- [28] Lin Chun-Mei, He Yue, Tang Bing-Yong. Application research in stock market forecast with fuzzy cognitive maps. *Chinese Journal of Computer Application*, 2006, 26(1): 195-201(in Chinese)  
(林春梅, 何跃, 汤兵勇. 模糊认知图在股票市场预测中的应用研究. *计算机应用*, 2006, 26(1): 195-201)
- [29] Zhang Qiang, Wang Xi, Liu Xiao-Dong. AFS theory, fuzzy cognitive map and data mining. *Chinese Journal of Dalian Maritime University*, 2004, 30(4): 106-109 (in Chinese)  
(张强, 王昕, 刘晓东. AFS 理论和数据挖掘与模糊认知图. *大连海事大学学报*, 2004, 30(4): 106-109)



**CHEN Zhuang**, born in 1964, Ph.D., professor. His research interests include enterprise informatization and knowledge management system.

**Montazemi Ali R.**, born 1951, Ph. D., professor. His research interests focus on intelligent information system.

## Background

This paper is partially supported by National Science Foundation of Chongqing under Grand No. 2005BB2083 entitled "Study on Management Technologies of Data Resources". The authors and their groups have done some works in areas of cognitive map (CM) and data resources management, such as CM' application for information requirements analysis, the relationship between CM and data mining, and CM' feedback mechanism. Recently, CM have been gained considerable research interest and applied to many areas, but the methods of obtaining the CMs, including questionnaire, brainstorming and sample learning method,

have some deficiencies, such as neglecting the objective data resources and losing information. Therefore, the paper focus on the problem of methodology of mining CMs based on data resources, which is significant step to exhibit CM' advantages and make it use in some practical domains. The authors' work is to create a new methodology of mining CMs from data resources. The paper firstly proposes the structure of mining CMs, then gives initialization technologies for the database and creates optimization algorithm for weight coefficients. At last, the authors design simplification strategy of CMs. The experimental results show the methodology is effective.