计算机科学 1998 Vol. 25 №. 2

企业智能支持系统的证据推理支持策略研究

On Evidence Reasoning Support Strategies of Intelligent Support System in Enterprises

孙 波 林宣雄 李怀祖

（西安交通大学管理学院 西安 710049）

摘 要 本文结合目前数据库、信息系统、专家系统分布化、网络化的趋势，讨论了在构造企业智能支持系统时的合作解题问题，研究了合作求解问题的典型类型，并对基于证据推理的相关支持技术进行了讨论。

关键词 人工智能，证据推理，企业智能支持系统

一、企业智能支持系统及其特点

随着计算机网络技术、数据仓库技术、数据挖掘技术、可视化计算技术的发展，企业信息系统的设计与开发均发生了重大的变化。企业智能支持系统是在企业内部网与企业内部信息系统的基础上为了更有效地协助企业各级用户完成各自的任务而开发设计的人工智能系统。企业智能支持系统的主要任务在于利用成熟有效的人工智能技术，结合用户具体解题任务的要求，自主地或在用户驱动之下，智能地完成有关任务并与用户交互。企业智能支持系统的基本结构层次如图 1 所示。

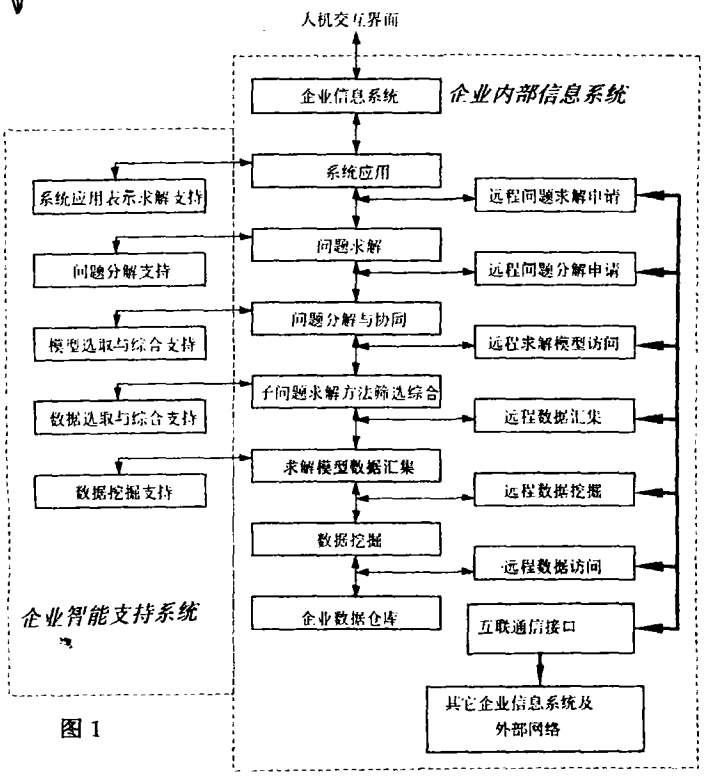

图 1

企业智能支持系统具有以下特点：

1）企业智能支持系统并非独立运行的一个完整系统，而是面向企业内部信息网的具体应用，由企业内部信息系统驱动后自主运行，或在用户监控与引导下被动运行；

2）企业智能支持系统是由支持不同层次企业内部信息系统应用的独立模块构成的分立系统，即为了保持企业内部信息系统的可扩展性与可剪裁性而牺牲自身的独立性；

3）企业智能支持系统应当保持一定的抽象性，以便保证在支持相近应用需求时可以提供一致的操作接口，从而提高自身的灵活性，降低复杂度；

4）企业智能支持系统应当可以处理模糊多语义的应用数据，即为同一组数据规定不同应用场合下的语义映射法则，从而提高有关应用模块的普适性和可扩展性；

5）企业智能支持系统的具体构成依赖于企业内部信息系统应用的具体组成。本文列举的层次模型

孙波 博士生。林宣雄 副教授、博士。李怀祖 教授、博士导师

分别支持企业内部信息系统的问题规划与表示、问题分解、求解模型选择、模型数据综合及数据信息挖掘。对于具体应用来说，未必严格遵循上述层次划分，但这种层次间的逻辑关系应当是企业智能支持系统的主要特点；

6)企业智能支持系统不追求求解结果的最优化和精确化，而应当追求普适、高效、灵活地解决企业内部信息系统的问题，尽可能地减轻用户的工作负担，动态地与有关信息源进行交互，并为用户提供有关问题的可行解或满意解；

7)企业智能支持系统的激活、运行、提交、终结的全过程必须是可控的，以利于用户依据具体需求来规定其运行方式，限制无限搜索，节约求解费用。

综合上述讨论，企业智能支持系统是由几组独立的智能模块组成的混合系统。本文研究的重点是与企业智能支持活动有关的各类模糊推理策略。从图1的企业智能支持系统层次模型可以看出，有关问题表示的支持技术及问题求解规划技术均依赖于有关应用系统的具体构成，而数据挖掘技术则是当前数据仓库与数据挖掘研究的主要问题，本文不拟对这两个问题进行深入细致的研究。本文研究的主要问题是：①依据企业智能支持系统的特点，如何选定合适的模糊推理支持策略，支持有关的测度合成问题；②在选定合适的模糊推理支持策略后，进行测度自学习的方法。

二、利用证据推理方法进行测度合成

本文要研究的问题，归根结底是测度合成问题。为此，在问题结构不清晰或缺少领域知识时，可以采用证据推理方法进行有关的智能支持工作。证据推理是根据证据理论得到的不确定性推理方法。它首先给出假设的度量函数 mass 函数，mass 函数可以由人们主观给定，或凭经验和感觉给出，由 mass 函数给出的信任测度和似然测度，作为对假设的信任测度估计的下限和上限。而信任测度和似然测度本身又可以得到条件信任测度与条件似然测度，进而得到证据推理模式。

利用证据推理方法进行企业智能支持系统设计的主要原因有：

1)证据推理中需要的数据更直观、更容易获得。在企业信息系统应用中，由于各类应用的灵活性较强、企业信息系统面向的外部环境瞬息万变，再加上使用者的主观因素必将以一定的形式反映到具体的应用结果之中，这一切都决定了企业智能支持系统运行中使用的测度依据是一组完整性与一致性较差，且变动不定的数据。企业信息系统的外部数据源的可管理性又很差，因而需要一种灵活性好，对原始数据要求不太高的推理方法；

2)证据理论在进行知识获取与加工时，并不要求证据与假设间有严格的因果关系，证据与假设间的关系是“伴生”的关系，其中关系的度量是 mass 函数。这样，一方面在企业内部信息系统应用中，人们不必深入了解和整理整个企业运行中涉及到的所有问题及其间的因果联系(这类工作往往耗费大量的企业资源)；另一方面，对 mass 函数的确定也不必全部交由专门的知识工程师来作，而可以由直接用户、领域专家、协作伙伴等共同给出。不同来源的 mass 函数可能千差万别，且其中涉及的“证据-假设”集合可以是异构的，这对于采集并表达不同知识源的知识是十分必要的；

3)证据推理中，可以通过 Dempster-Shafer 合成公式综合不同知识源的知识。在应用中，这也是修正不良知识源知识的重要手段；

4)证据推理方法可以在证据采集不完备的情况下工作。当推理所需的有关数据采集不完备时，系统仍可以推理，并将有关未采集信息从推理过程排除出去，并通过假设估计上下限的变化来表征出来。这对于分布运行、更新迅速、外联广泛的企业信息系统是十分重要的；

5)通过对有关机器学习和人工智能解决方案的研究，进一步讨论有关基准测度生成方法、测度更新与管理及知识发现策略，可以改善企业内部信息系统的工作状况，提高其智能化水平。

当证据或者出现或者不出现时(对应于企业智能支持系统应用，即测度依据是对某些命题的是非判断时)，可以通过证据推理原理进行测度合成。在企业智能支持系统工作中的有关算法是：

1. 确定假设空间 H 及证据空间 E；

2. 依据以往的经验及信息系统运行记录，确定假设空间的先验概率 P；

3. 依据系统收集的有关测度知识，利用合成公式，生成与推理有关的 mass 函数；

4. 随机地向部分信息源发出数据采集请求；

5. 依据搜集的数据，确定证据的概率分布；

6. 利用证据推理原理，计算假设的信任区间；

7. 进一步随机地向其他信息源发出数据采集请求，如果返回信息对假设信任区间的修正作用大于预定的阈值，则转向5；

8. 否则，提交测度生成结果。

在测度生成中，可以由测度生成模块的上级调用者依据求解的要求及求解资源约束，确定某次特定求解过程的阈值条件，从而提高有关智能支持活动的可控性。

当证据本身是不确定的(对应于本文的研究，即企业智能支持系统进行测度合成的指标依据是一组模糊数据时)，可以采用证据推理模式进行测度合成。其过程类似于证据推理原理在指标合成中的有关算法，这里就不详细论述了。

以上我们讨论了证据推理方法在问题结构不清晰且缺乏领域知识时对测度生成的支持作用。当具备足够的领域知识时，我们可以采用启发式规则来进行测度估计。当问题结构化程度高且存在对待求解问题的完整描述时，可以通过模糊综合评价方法进行综合测度评价。对于结构固定且输入输出间存在隐函数映射关系的场合，可以采用神经网络方法来学习隐含规则并进行工作。由于这些方法都是相当成熟的技术，限于篇幅这里就不一一讨论了。

三、测度函数自动生成问题研究

在证据推理的研究中，有不少学者讨论了利用关系数据库进行证据推理的有关方法。在研究中，可以利用关系数据库记录表示外延空间与内涵空间，建立外延空间的概率分布，确定内涵空间到外延空间的映射或关系划分。关系划分的一种直观形式是，对于一个内涵空间的子集 a，其关系划分为外延空间中与 a 中所有内涵均有关的元素的集合 d(a)。证据理论证明，d(a)出现的概率即是内涵空间子集 a 的 mass 函数。这样就可以直接建立对内涵空间的测度。依据关系数据库的证明推理方法，如果我们视外延空间为证据，内涵空间为假设，可以根据关系数据库中的信息，通过人为给定的 mass 函数构造方法得到对假设空间的 mass 函数。对于给定的一组特定证据，我们可以计算假设空间中信任测度最大的假设子集。作为证据推理的结论，并由此结论计算信任区间。这构成了一种可行的利用案例信息进行证据推理的方法。

利用这类方法进行证据推理从理论上讲是可行的，也是有效的。但对于我们的系统要求来说，不仅需要一种理论上有效可行的方法，更重要的是建立一种能尽可能准确的证据-假设生成机制，换句话说，我们要从所有 mass 函数的可能构造中筛选出能尽可能准确地由证据生成假设的一种。关系数据库的证据推理方法提供了与证据推理理论本身自满足的许多可行方法，它们具有形式规范、简单等优点，但不一定满足我们进行推理的要求。为此有必要讨论一种满足我们的案例推理系统要求的测度函数自学习求解方法。以下，我们以证据推理原理中的 mass 函数自学习为例进行研究

3.1 测度函数自学习样本集的确定

通过自学习算法求得的 mass 函数只是依据已有的案例样本得出的一个较有效的测度函数，但对于更多的未知的样本是否有效则很难确定。如果存在专家给出的其它 mass 函数，我们在求解未知样本时应综合它们的结果。如果案例集中证据与假设的关系较简单，案例集较充分，且自学习算法得出的 mass 函数比综合后的 mass 函数更好，那么可考虑用系统求解出的 mass 函数单独工作。

为了能够自学习地求解 mass 函数，我们需要一个用于学习的样本集合，样本集合中的元素(即样本)可以表达为：$(f(e_1),f(e_2),\cdots f(e_n),g(h_1),g(h_2),\cdots,g(h_m))$。其中 $f(e_i)(1\leqslant i\leqslant n)$表示对证据 e_i 的度量，$g(h_j)(1\leqslant j\leqslant m)$表示对假设 h_j 的度量。对于证据推理原理来说，$f(e_i)$表示 e_i 的概率分布。$g(h_j)$表示 h_j 的信任区间。对于证据推理模式来说，$f(e_i)$表示 e_i 的信任区间。$g(h_j)$表示 h_j 的信任区间。

样本集筛选的目的是确定与一个特定假设集合相关的证据集合，在算法设计时，我们以某一特定待测指标 A 的不同取值 $a_1,a_2,\cdots,a_m$ 作为假设集合，对于 A 取连续值的情况，应先对指标 A 作离散化。至于离散的方法可以由系统指定，原则是在尽可能降低离散化程度的前提下，保持一定的模糊匹配精度。一种直观的方法是先由专家指定离散化程度，并对指标进行离散化。然后对系统记录的应用，即用户使用的情况进行测试。如果发生匹配精度下降的情形，例如与用户需求匹配的案例太多，则可以试着将某个离散化级别作分解。计算该分解提高匹配精度的程度。保留匹配精度，提高最大的分解。重复这一步骤，直至匹配精度达到要求为止。

对于证据推理原理来说，若 f(e)在两个不同样本中取值为 $f(e_1),f(e_2)$，则当 $|f(e_1)-f(e_2)|>t$ 时，认为 e 发生了变化。对 g(h)的不同样本取值 $g(h_1)$，$g(h_2)$，当其相似度 $r(g(h_1),g(h_2))\leqslant 1-s$ 时，认为 h 发生了变化。

对于证据推理模式来说，若 $r(f(h_1),f(h_2))\leqslant 1-t$ 时，认为 e 发生了变化。$r(g(h_1),g(h_2))\leqslant 1-s$ 时，则认为 e 发生了变化。相似度的计算可参见前面关于模糊匹配的论述。我们称证据集是不完备的，是指对于样本集中某个假设 h 的变化，无任何相应的

证据变化。

这样，我们可以得到一个特定的假设空间 $H=\{h_1, h_2, \cdots, h_n\}$，确定其相关证据空间 E 的方法如下：

①选取全部证据的经验集合 $X=\{x_1, x_2, \cdots, x_l\}$，其中 $x_i(1\leqslant i\leqslant l)$为曾在某个案例记录中出现的证据。

②删除 X 中的无关证据。取 X 的子集 $Y=\{y_1, y_2, \cdots, y_k\}$，其中，对于任意 $y_i(1\leqslant i\leqslant k)$存在 $h\in H$，并有案例记录说明 y_i 与 h 有关。

如果不考虑证据与假设之间有无逻辑上的联系，为了减少与假设集相关的证据集的规模，还可以进行如下操作：

③删除 Y 中的伴生证据，对于 Y 中的两个证据 y_1, y_2，如果二者的变化总是同时发生，则可删除其中任何一个。

经过以上运算，得到的证据集合 Y 就是所求的证据空间 E。将证据空间和假设空间 H 合并可得到所需的样本框架 $(e_1, e_2, \cdots, e_n, h_1, h_2, \cdots, h_m)$；其中 $\{e_i\}=E$；$\{h_i\}=H$。再依据该框架从案例集中抽取有关样本 $C_j=\{f_j(e_1), f_j(e_2), \cdots, f_j(e_n), g_j(h_1), g_j(h_2), \cdots, g_j(h_n)\}$即为第 j 个样本。所有样本的集合 T 即自学习样本集。

3.2 基于证据推理原理的自学习算法

自学习算法求解 mass 函数的目的是使得利用该 mass 函数进行推理时，由证据推理得到的假设尽可能精确。为了达到这一目的。必须先确定精确度的度量标准。

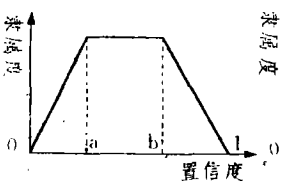

图 2

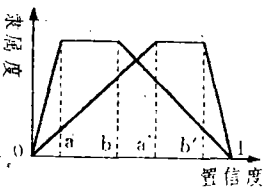

图 3

如图 2，若假设 h 的置信区间为 $[a,b]$$(0\leqslant a\leqslant b\leqslant 1)$，则我们可以定义假设 h 的置信度的隶属度函数：

$$h(x)=\begin{cases} x/a; & 0\leqslant x<a \\ 1; & a\leqslant x\leqslant b \\ (x-1)/(b-1); & b<x\leqslant 1 \end{cases}$$

如图 3，对于假设 h 及 h′，若其置信区间分别为 $[a,b]$及$[a',b']$，可依上式定义置信度的隶属度函数：h(x)和 h′(x)。我们定义其相似度为：

$$R(h,h')=\frac{\int_0^1 \min\{h(x),h'(x)\}}{\int_0^1 \max\{h(x),h'(x)\}}$$

对于任意一个样本 $C=(f(e_1), f(e_2), \cdots, f(e_n), g(h_1), g(h_2), \cdots, g(h_m))$，若由系统根据同样证据条件推得的期望样例为：$C'=(f(e_1), f(e_2), \cdots, f(e_n), g'(h_1), g'(h_2), \cdots, g'(h_m))$，则可如下定义样本 C 的求解精度 S(C)：

①最大相似度原则。即：$S(C)=\max\{R(h_i, h'_i)\}$；$(1\leqslant i\leqslant m)$

②加权平均原则。即：$S(C)=\sum_{i=1}^{m} w_i\times R(h_i, h'_i)$。其中 $\sum_{i=1}^{m} w_i=1$；且 $0\leqslant w_i\leqslant 1$；$(1\leqslant i\leqslant m)$

这样对于全体样本集 T，设 $T=\{C_1, C_2, \cdots, C_l\}$，相应可定义样本集求解精度 S(T)：

①最小样本求解精度原则。即：$S(T)=\min\{S(C_i)\}$；$(1\leqslant i\leqslant l)$，其中 $S(C_i)$依据最大相似度原则求出；

②算术平均原则。即：$S(T)=\frac{1}{l}\sum_{i=1}^{l} S(C_i)$；其中 $S(C_i)$依据加权平均原则求出。

针对指标 H，可根据 H 的可能取值，建立假设空间 $D=\{h_1, h_2, \cdots, h_m\}$，其中，$h_i(1\leqslant i\leqslant m)$为 H 的第 i 个取值等级。若此时，与 H 取值相关的证据空间为：$E=\{e_1, e_2, \cdots, e_n\}$，$e_i(1\leqslant i\leqslant n)$为 H 的第 i 个证据。可建立自学习样本集 $T=\{C_1, C_3, \cdots, C_n\}$，其中 $C_i=\{f_i(e_1), f_i(e_2), \cdots, f_i(e_n), g_i(h_1), g_i(h_2), \cdots g_i(h_m)\}$为从案例集抽样得到的样本。根据样本集统计，可得假设空间的统计频度，当样本集足够大时，可以用频度逼近 D 的先验概率。我们在算法中，以统计频度作为先验概率的估计值，可得 $P=\{p(h_1), p(h_2), \cdots, p(h_m)\}$。

自学习算法的要点是通过自学习过程，建立对于每个 e_j 的 mass 函数 $m(./e_j)$及 $m(./\neg e_j)$。在设计算法时，首先要确定 mass 函数的论域。从前面对证据推理的讨论中可以看出，mass 函数的论域为整个 D 空间上全体子集的集合 A。如果我们在自学习算法中，也将待学习的 mass 函数论域定为集合 A，那么由于集合 A 中元素的个数为 2^n，其中 n 为 D 中元素的个数。这样，算法的空间复杂度会随着 D 的增加以指数速度增加，为解决这个问题，须对 A 进行简化。

从我们对 D 的定义可以看出，对于 D 中两个元素 h_1, h_2，如果 $h_1\neq h_2$，则 h_1, h_2 不应是伴生假设。即当 h_1 成立时，h_2 一般不成立。这样我们将 A 中除元素 D 外，其余包含两个以上元素的子集舍去，或设其 mass 测度为 0，理由是：

①因为包含两个以上元素的子集在证据推理中几乎不成立，设有一个包含两个元素的集合 $B=\{h_1,h_2\}$，若 $mass(B/e_j)\neq 0$，如果证据推理是完备的，必然存在一种证据组合，在这个组合下，证据推理的结论为B，即两个互斥的测度成立。这对于样本集较小、训练程度不高的情况尚有合理性，对于样本充分、训练程度高的情况则不合理。

②如果对某个证据 e_j 来说，存在A中的一个包含两个以上元素的子集 $B=\{h_1,h_2\cdots\}$，若 $mass(B/e_j)>0$，则说明 e_j 的出现会使若干相互矛盾的测度成为可能，这在证据理论研究中是有意义的，但对于自学习求解来说，我们可以认为导致这一情况的原因是 e_j 并非单纯的证据，而是一个复合证据，对 e_j 可以进行分析和分解，使得对于任何 $B\subset A$，若 $|B|\geqslant 2$，则 $mass(B/.)=0$。

需要说明的是，上述假定只有在不严重影响mass函数可信度的情况下对算法进行空间复杂度优化才具意义。在证据推理原理中，未必要求两个以上假设的mass测度为0，在假设空间较小时，仍可选择D中子集的全集A为mass函数的论域。下面给出基于证据推理原理的mass函数自学习算法：

步骤一：确定样本集 $T=\{C_1,C_2,\cdots,C_l\}$。其中 $C_i=(f_i(e_1),f_i(e_2),\cdots,f_i(e_n),g_i(h_1),g_i(h_2),\cdots,g_i(h_m))(1\leqslant i\leqslant l)$，假设的先验概率 $P=\{p(h_1),p(h_2),\cdots,p(h_m)\}$。确定待训练mass函数的论域A（当A中元素较多时，$A=\{\varphi,\{h_1\},\{h_2\},\cdots,\{h_m\}D\}$）；

步骤二：对于A中任意元素 $a\in A$ 及任一证据 $e\in E$，设定mass函数初值：$m(a/e_j)$ 及 $m(a/\neg e_j)$，使：a) $m(\varphi/e_j)=m(\varphi/\neg e_j)=0$；b) $m(a/e_j)\in[0,1]$，$m(a/\neg e_j)\in[0,1]$；c) $\sum_{a\subset A}m(a/e_j)=1$；$\sum_{a\subset A}m(a/\overline{e_j})=1$；

在实际计算时，可以令 $m(\varphi/e_j)=m(\varphi/\neg e_j)=0$；对 $a\in A$ 且 $a\neq\varphi$，令

$m(a/e_j)=m(a/\overline{e_j})=\frac{1}{|A|-1}$；其中 $|A|$ 表示A中元素的个数；

步骤三：确定自学习步长d。$d\geqslant 0$，为一个很小的正数（$\ll 1$）

步骤四：对于全体样本集，计算样本求解精度 $S(T)$（详见《不确定性推理原理》）

步骤五：对于 $e_j\in D$，选择 $a\in A$，在调整操作不违反 $m(./e_j)\in[0,1]$ 的前提下：

a)若令 $m(a/e_j)=m(a/e_j)$原始值$+d$，对于其它 $b\in A$，令 $m(b/e_j)=m(b/e_j)$原始值$-d^*$[$m(b/e_j)$原始值]/[$1-m(a/e_j)$原始值]；计算此时样本集求解精度 $S_aT(j,a)$；

b)若令 $m(a/\neg e_j)=m(a/\neg e_j)$原始值$+d$，对于其它 $b\in A$，令 $m(b/\neg e_j)=m(b/\neg e_j)$原始值$-d^*$[$m(b/\neg e_j)$原始值]/[$1-m(a/\neg e_j)$原始值]；计算此时样本集求解精度 $S_bT(j,a)$；

c)若令 $m(a/e_j)=m(a/e_j)$原始值$-d$，对于其它 $b\in A$，令 $m(b/e_j)=m(b/e_j)$原始值$+d^*$[$m(b/e_j)$原始值]/[$1-m(a/e_j)$原始值]；计算此时样本集求解精度 $S_cT(j,a)$；

d)若令 $m(a/\neg e_j)=m(a/\neg e_j)$原始值$-d$，对于其它 $b\in A$，令 $m(b/\neg e_j)=m(b/\neg e_j)$原始值$+d^*$[$m(b/\neg e_j)$原始值]/[$1-m(a/\neg e_j)$原始值]；计算此时样本集求解精度 $S_dT(j,a)$；

求 $S_fT(J,a)=\max(S_aT(j,a),S_bT(j,a),S_cT(j,a),S_dT(j,a))(1\leqslant j\leqslant m,a\in A)$。若 $S_fT(j,a)\leqslant S(T)$，转步骤七；

步骤六：依 $S_fT(j,a)$，对 $e_j\in E$，$a\in A$，按步骤五中的调整方法f进行调整，令 $S(T)=S_fT(j,a)$，转步骤五；

步骤七：这时的mass函数取值就是对论域A进行证据推理的较精确的测度函数，记录各mass函数值并输出。

四、有关问题的讨论

从企业智能支持系统实现及系统运行的技术要求上讲，上面讨论的证据推理方法适合于处理与某特定指标测度相关的证据明确、测度简单、主要证据可以枚举的场合。最理想的情况是证据测度方法规范、可重复操作、同样环境下证据测度可重现。而且证据与待测指标间满足映射关系，即相同证据测度下，不应导致截然不同的指标测度。

如果以待估计指标作为证据出现的诱因，则证据可视为待测对象的变化作用于外界环境（或系统自身）后表现出来的症状。从这种意义上讲，证据推理是"执果导因"的操作，由于不存在其它制导信息，因此推理结果如何将严重依赖于证据集合的确定。在选取证据时，应在待测对象导致证据的因果链上选取与诱因尽可能接近的证据。同样，如果证据是引起对象性质变化的其它原因的外推表现，则在选取证据时，应坚持两个原则：一是选择与被测对象变化尽可能直接相关的变化诱因；二是选择诱因的尽可能直接的证据表现。

为了使系统能识别案例记录中与待测对象性质无直接关系的证据，我们可以通过以下步骤筛选并淘汰这类证据：

步骤一：建立所有待测指标的相关证据集合的并集。凡不在该集合中的证据与任何待测指标均无关，可以舍弃；

步骤二：a)对于证据推理原理自学习算法生成的mass函数，设有阈值 $T(0\leqslant T\leqslant 1)$，若 $\max(m(./e_j))\leqslant T$ 且 $\max(m(./\neg e_j))\leqslant T$；即无论 e_j 是否出现，对推理结果都无重大影响，则认为 e_j 属无效证据，可以剔除。b)对于证据推理模式自学习算法生成的规则"$e_i\rightarrow h_j$"的信任区间$[B(i,j),L(i,j)]$，设有阈值 $T(0\leqslant T\leqslant 1)$，若 $\max(L(i,j))\leqslant T(h_j\in D)$，则一定有 $\max(B(i,j))\leqslant T$；即 e_j 任何假设都无重大预见性，则认为 e_j 属劣质证据，可以剔除。

步骤三：若一个证据可以由其它证据推得，则该证据属冗余证据，应删除。这时可将待删除证据作为假设，其它证据作为求解该假设的依据。设有阈值 $T(0\leqslant T\leqslant 1)$，若依自学习算法求得待删除证据的测试样本集求解精度 $S(T)>T$，则删除该证据。

为了保证证据推理的有效性，经筛选的证据集最好是完备的。所谓完备是指通过现有的证据集合可以确定待测假设集合。如果满足这一条件，则证据推理方法就可视为完全可信的假设集求解办法。如果证据集不完备，就应提醒系统有进一步搜集证据的必要，判断一个证据集是否完备，对于提高系统的自学习水平和提供尽可能好的指标估计是十分必要的。我们可以根据下面两个简单原则来发现一些证据集不完备的情形。

原则一(无冲突原则)：任何对于假设进行测度的完备集均应满足在相同的证据条件下不会出现冲突的假设采样。我们可以从对某个假设测度的样本集中选出所有相同证据条件的样本，若这些样本中对假设的案例采样有较大差异(可设定一可行的阈值来判断)则认为证据集不完备，须采集对该假设测度的其它证据。

原则二(推理成功率下限原则)：经过样本集训练得到的证据推理方法应满足一定的推理成功率下限。对于一个特殊的例子，设证据集为$(e_1,e_2,\cdots,e_m)$，$e_i(1\leqslant i\leqslant m)$的取值为"出现"或"不出现"，即为二值变量。则此证据集的样本集最大样本个数为 2^m，此时有：

(1)若假设集可构成的样本集最大样本个数超过 2^m，则证据集不完备。

(2)若已由 k 个不同训练样本进行自学习，则理想情况下，自学习样本可以经证据推理重演，即证据推理成功率下限为 $k/2^m$。如果对测试样本集的测试命中率$<k/2^m$，则可认为证据集不完备。

(3)还可以由系统得出对 k 个训练样本自学习后的推理命中率下限，做为判别完备性的依据。

参考文献

[1] 傅京孙、蔡自兴、徐光，人工智能及应用，清华大学出版社，1987

[2] 张文修，不确定性推理原理，西安交通大学出版社，1994

[3] 张文修、陈雁，合情推理与发现逻辑，贵州科技出版社，1994

[4] 孙波、袁慧萍、李怀祖，管理专家系统与情境研究，计算机科学，6(23)1996

(上接第 50 页)

0.25～0.75。P_m 太大将产生过多的预测模型串，P_m 太小又会导致不产生新的预测模型串，建议取值 0.01～0.20。

结束语 本文研究了通过遗传算法寻找最佳的预测模型的方法。笔者仅以环境质量预测为例，对 Brown、Horton、Prati、Nemerow 等大气评价模型及有关数据进行二进制编码后，使用本遗传算法寻找到不同污染情形下的预测模型，具有相当的适用价值。经过研究笔者认为遗传算法特别适合于结构复杂的非线性问题，并能对大多数的问题给出满意的解，而且比其它算法更容易实现。但如何将问题进行二进制编码值得进一步研究，这将直接影响问题求解的精度和遗传算法收敛的速度。

参考文献

[1] 张晓缋等，一种新的优化搜索算法—遗传算法，控制理论及应用，22(3)1996

[2] 韩桢祥等，模拟进化优化方法及应用，计算机科学，22(2)1995

[3] 王丽微等，遗传算法的收敛性研究，计算机学报，19(10)1996

[4] 张玲等，统计遗传算法，软件学报，8(5)1997

[5] 陈恩红等，基于遗传算法的概念学习中的约束满足预处理方法，计算机研究与发展，34(7)1997