

应用面向对象知识处理的选案专家系统

谭 宁 (中国科技大学研究生院 100039)
叶施仁 游湘涛 史忠植 (中科院计算技术研究所 100080)

摘要:本文介绍了一种用专家系统实现税务部门计算机稽查选案的方法。这一计算机选案专家系统是用面向对象的知识处理系统(OKPS)构造而成。OKPS是我们自主开发的一套可视化专家系统工具。本文着重阐述了OKPS系统和该专家系统的原理、选案知识库的构造、推理控制策略等问题。本文所讲解的这一计算机选案专家系统已作为广东地税信息系统(GDLTIS)的一部分投入实际运行。

关键词:专家系统 知识处理 面向对象 税务稽查 计算机选案

一、计算机选案系统的介绍

1. 什么是计算机选案

国家税务总局的税务稽查业务包括了选案、实施、审理和执行四个环节。

选案的方法主要有人工选案和计算机选案两种。

人工选案即利用现有的信息数据资料,通过人工分析、归集分类从而确定稽查对象。

计算机选案,是指依据科学的选案标准,利用计算机对现有的数据资料进行加工处理,分析筛选,发现对税收影响异常的纳税人。通过计算机选案选出的可疑户,经过适当筛选后将发送至案源库,供稽查计划环节制定稽查计划时使用。

计算机选案可以在一定程度上克服人工选案的任意性和盲目性,它尤其适合于大数据量情形。当然,在计算机选案推行的初期,由于选案指标或参数不一定很合理、或者由于基础数据不够准确完备,这都会影响到计算机选案效率的发挥。

随着税收电子化的逐步深入,纳税人数据的采集将会逐步完备和准确;随着选案手段的逐步科学化,越来越多的选案经验会转化为选案指标,选案系统会逐步丰富和成熟。

本文所要讲解的计算机选案专家系统是由中科院计算所的智能计算机科学实验室与广东地税局合作开发完成,是广东地税信息系统(GDLTIS)的一部分。

2. 计算机选案的基本原理和有关概念

计算机选案(以下简称“选案”)是指通过计算机按预定规则(条件)对纳税人的数据进行加工、判断、分析,找出对税收影响异常的可疑户。在这一过程中,计算机还

将基于一定规则对可疑户进行打分排序,产生按分值排序的选案清册表和选案指标分析表,以及其他的决策和解释信息,供计划环节制定稽查计划时使用。

选案的数据来源包括纳税人在办税过程中申报的、税务机关受理并审核后分别于各个不同的工作环节录入计算机信息库的有关数据资料,纳税人财务报表数据,以及审理环节录入计算机的各种稽查结果数据资料。例如,税务登记资料、发票管理资料、减免税申请资料、纳税申报资料和稽查结果资料等。除此以外,还可能包括从外部采集得到的工商信息、房产信息、交通管理部门信息等。为了讨论的方便,我们把它们称为选案变量。

选案指标是从某一个方面评价纳税是否异常的标准。它根据与该纳税人有关的数据(选案变量),按照给定的规则和方法,从某个方面评价是否纳税异常。例如,从销售税金负担率的方面考虑,如果现在的销售税金负担率比上一个时期(上期)的低,或者比同行业的低,则可能纳税异常。又如,利润是忽盈忽亏的,则可能纳税异常。选案指标涉及两个问题:一个是选案变量从何处抽取数据?另一个是怎样使用这些选案变量?不同的指标完全可能用到同样的选案变量,因此我们将前者集中起来讨论和研究。在选案指标中只关心怎样使用这些选案变量的规则和方法。选案指标描述了由选案变量得到指标分数的规则。此外,为了量化纳税异常的程度,对每个指标给出一个分数。这个分数叫指标分数。为了问题的简单,我们规定,指标分数越高,纳税异常可能性越大,纳税异常程度越高。指标分数为0,则表明这项指标正常。

选案方法由若干个选案指标组成,用户可以使用它进行计算机选案。如果选案方法仅由一个选案指标组成,则称之为单项指标打分方法,否则称为综合打分方

法。一个选案方法中所有选案指标用到的选案变量组成该选案方法的选案变量。选案方法同样由一个反映纳税异常程度的分数(总分),它的大小由指标分数决定。总分并非指标分数的算术平均值或者代数和,所有的选案指标并非重要程度都一样,他们通过权值体现出来。总分越高,排名越靠前,纳税异常可能性越大,纳税异常程度可能越高。如果总分为0,则表明用这种方法进行计算机选案时,纳税情况正常。

下面我们来举例说明这些概念。对“销售利润率”这个指标,其目的是判断某纳税人现在的销售利润率(本期)是否比以前的(上期)低,或者(和)比同行业的低,如果是这样,它可能隐瞒收入,因而可能税收异常。它包括:

·计算公式:

本期销售利润率 $x_1 = \text{本期销售利润} / \text{本期销售收入}$;

上期销售利润率 $x_2 = \text{上期销售利润} / \text{上期销售收入}$;

同行业销售利润率 $x_3 = \text{同行业销售利润} / \text{同行业销售收入}$;

·打分规则:

设 $k_1 = x_1 - x_2$ 。如果 k_1 小于0,则 k_1 越大,指标分数越高,否则指标分数为0;

设 $k_2 = x_1 - x_3$ 。如果 k_2 小于0,则 k_2 越大,指标分数越高,否则指标分数为0。

它使用的选案变量为:本期销售利润、上期销售利润、同行业销售利润、本期销售收入、上期销售收入、同行业销售收入。

对“销售税金与销售收入变化幅度(同步法)”这个指标,其目的是判断销售税金与销售收入是否同步变化,如果不同步,则可能纳税异常。它包括:

·计算公式:

销售税金变化幅度 $x_1 = (\text{本期销售税金} - \text{上期销售税金}) / \text{上期销售税金}$;

销售收入变化幅度 $x_2 = (\text{本期销售收入} - \text{上期销售收入}) / \text{上期销售收入}$ 。

·打分规则:

如果 x_1 与 x_2 的符号不同,则可以根据 x_1 、 x_2 差异给指标打分,否则指标分数为0。

它使用的选案变量为:本期销售税金、上期销售税金、本期销售收入、上期销售收入。

这两个选案指标可以分别构成两个单项指标的选案

方法,也可以将它们以及其他指标组合起来形成一个综合打分方法。

由以上说明可以看出,为了准确而有效地进行选案,系统必须要做到以下几点:

- 能取得有效而一致的数据;
- 制定合理的判断和分析的规则条件;
- 具有灵活高效的控制策略,用来指导和完成选案过程;

·提供丰富且必要的分析和解释信息,使用户能够了解计算机选案的分析判断过程和依据。

·具有良好的灵活性、易操作性和可扩充性,能够便捷的产生、修改和维护各种选案指标和选案方法。

这些要求和特点表明,计算机选案适合用专家系统来解决。为了高效地构造出功能足够强大的计算机选案专家系统,我们应用了面向对象知识处理系统(OKPS),一个我们自主开发的可视化专家系统工具。下面来简单介绍一下这一专家系统开发工具。

二、面向对象的知识处理系统(OKPS)

这一专家系统工具由两部分组成:知识获取和管理工具,以及面向对象推理机。知识获取和管理工具是一个可视化的知识库构造和管理程序。用这个工具构造的知识库,可以在面向对象推理机中访问并进行推理。系统的主要功能和特点是:

1. 采用面向对象的知识表示方法

知识表示是指用计算机表示知识的可行的、有效的、通用的原则和方法。该系统将框架理论和语义网络相结合,采用面向对象的概念和技术来实现知识表示。一个知识库是由各种对象(Object)组成的树型结构。这些对象对应于树中的结点,它们有定义自身外观的属性和定义自身行为的方法。通过设置推理结点的属性和编写实现行为的方法脚本,可以将推理的规则和控制策略与对象有机的结合在一起。

2. 可视化的知识获取和管理

知识库的实现建立在技术成熟的数据库系统之上,通过通用接口,知识库可以建立在任何流行的数据库系统之上。通过专家系统工具中的可视化知识获取和管理工具,知识工程师可以方便地添加、删除和修改专家知识,从而高效地建立、扩展和维护专家系统的知识库。

3. 推理控制语言 ICL

该系统采用知识库系统和推理机制相分离的方式来构造专家系统。这样,构造专家系统时,只需要改变知识

库系统中的知识,就可实现不同的专家系统。为了提供功能足够强大的推理机制,并最大限度地保持灵活性和方便性,系统引入了一种专用的高级语言:推理控制语言 ICL,用来描述和操作专家知识与规则、以及控制推理过程。

ICL 脚本语言的元素包括构成 ICL 脚本的名字、数字、和字符。ICL 定义了如下一些符元(Token):标志、注释、关键字、标识符和常量等。

ICL 语句控制着脚本执行流程、象其他的编程语言一样在,ICL 中有几种语句可以执行循环,选择执行其他语句或是改变控制,这些语句包括:复合语句、do-while 语句、表达式语句、for 语句、if-else 语句、空(null)语句、返回(return)语句、规则(rule)语句和 while 语句等。

用户还可以使用功能强大而丰富的 ICL 函数。这些函数包括:控制台输入/输出、数值计算、字符串操作、图形/图表展示、黑板存取、文件操作、数据库访问、网络通信、知识库存取、推理控制以及外部调用等等。

ICL 中的数据共有 4 种类型:整型(int),浮点型(float),字符型(char),字符串型(string)。

4. 面向对象推理机制

在面向对象推理机中,知识库中每个对象中的方法将按一定顺序执行。每一个对象可以有三个方法:先序方法,中序方法和后序方法。通过知识获取和管理工具,你可以在这三种方法中编写推理控制语言(ICL)脚本。

在面向对象推理机中,推理过程是对知识库树型结构的遍历。在默认控制下,这个过程从根结点开始,树中所有结点都将被访问。

三、选案方法专家系统

1. 系统概述

计算机选案专家系统包括推理机和选案方法维护工具两部分,这两部分分别由 OKPS 的知识获取和管理工具以及面向对象推理机定制而来。选案方法维护工具用来创建、删除和修改选案方法知识库;推理机的运作就是根据选案方法知识库,分析加工纳税人的信息,进行综合打分的过程。

其原理是:系统根据用户选定待查的选案时间和范围、使用的选案方法,从数据库中取得这些纳税人的基本信息、财务报表和相关的征缴信息,并加以整理得出有用的选案变量;然后使用选案方法知识库中的知识,分析并计算出每个选案指标的分数以及该选案方法的总分。最后,系统给出按照总分的排名、有关的分户清册和选案过

程的解释说明。

本系统有两种工作状态:选案运行和维护。选案运行是计算机根据选案方法打分的过程,维护是指维护保存了选案方法知识库和选案方法用到的选案变量。前者是日常运行的,后者由熟悉系统的高级用户完成,不一定经常运行。

选案打分运行的过程如下:

- (1)确定选案范围、时间、行业以及所用的选案方法;
- (2)计算机自动生成所需要的选案变量;
- (3)推理机执行选案打分;
- (4)保存和查看选案结果。

维护的过程如下(以生成新的方法为例):

- (1)确定新方法的名称、涉及的税种、所用的选案变量等信息;
- (2)定义或修改它用到的选案变量的数据来源和加工方法;
- (3)通过选案方法维护工具,增加和修改指标;
- (4)试运行选案打分的过程,根据提示的信息重复以上步骤修改。

2. 选案方法知识库

选案方法知识库由一个叫选案方法的根结点和若干个叫选案指标的子结点组成,其中选案指标接点还可以包含其他子指标结点。整个选案方法知识库是一棵层次结构的树。

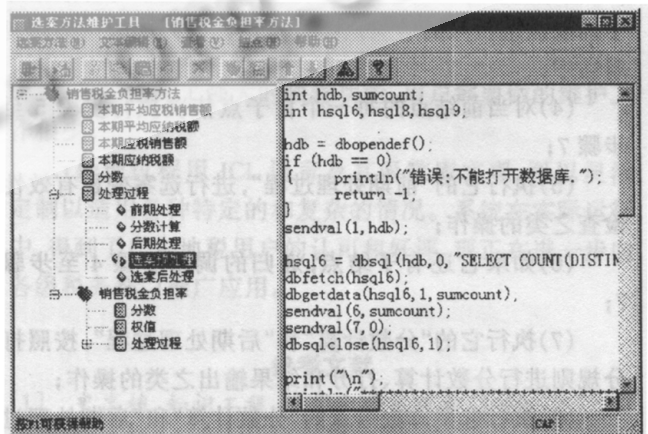


图 1 选案方法知识库(示意图)

在左边的用户窗口中,显示了选案方法知识库的结构。每一个选案方法以树的形式保存在知识库中,根结

点的名字就是选案方法的名称,它包括的子结点与这个方法的选案指标一一对应。根结点称为选案方法结点,子结点称为指标结点,指标结点还可以包含子结点(子指标结点)。结点本身包括两个方面的内容:属性和处理过程(相当于方法)。属性是用来保存数据的地方,保存的值分为整数、浮点数和字符串3种类型。用户可以指定它的缺省值,也可以通过 ICL 函数来访问和修改它。缺省情况下,对选案方法结点,它的属性记录了选案所需要的选案变量和总分;对指标结点,它的属性记录了指标分数和权值。处理过程描述了怎样使用选案变量。对指标结点,包括前期处理、分数处理和后期处理3个处理过程;对选案方法结点,还包括选案预处理、选案后处理2个处理过程。处理过程中包含的是 ICL 脚本。由于用户可以自由的编写脚本,所以具有很大的灵活性,能够方便地表示各种打分规则和计算公式,也能够处理各种复杂的情况。

事实上,系统会为结点生成缺省的脚本,一般情况下,用户不必对其做什么改动。针对每个指标结点,用户在构造选案方法知识库时,要做的工作主要集中在“分数计算”处理过程中。

专家系统在运行选案时,这些处理过程在缺省情况下是按照下列次序执行的:

(1)执行选案方法结点的“选案预处理过程”,为开始选案做打开数据库之类的准备工作;

(2)对每一个纳税人重复步骤3至步骤8;

(3)执行选案方法结点的“前期处理过程”,有数据库中抽取该纳税人的有关选案变量值读到方法结点的相应属性中;

(4)对当前结点的每一个子点依次执行步骤5至步骤7;

(5)执行它的“前期处理过程”,进行选案变量有效性检查之类的操作;

(6)如果它还有子结点,递归的调用步骤4至步骤7;

(7)执行它的“分数计算”和“后期处理过程”,按照打分规则进行分数计算、打分和结果输出之类的操作;

(8)执行选案方法结点的“分数计算”和“后期处理过程”,进行总分计算和结果输出之类的操作;

(9)执行选案方法结点的“选案后处理过程”,进行关闭数据库之类的操作;

(10)完毕;

3. 结点的处理过程

(1)选案方法结点的处理过程。在选案方法结点中,列出了用户选定的选案变量,它们是以方法结点属性的形式出现的,除此以外还包括一个叫“分数”的结点用来存储总分。选案方法结点包括5个处理过程,这个处理过程中包括了系统自动创建的脚本,这些缺省的脚本在一般情况下能够满足用户的需要,但必要时是可以作相应的调整。

·选案方法结点选案预处理过程的主要任务是打开数据库中相关的数据表。

·选案方法结点前期处理过程的主要任务是抽取当前纳税人的选案变量。

·选案方法结点分数计算过程的主要任务是根据制定的规则和公式计算选案总分。

·选案方法结点后期处理过程的主要任务是将结果输出到相应的选案结果表中。

·选案方法结点选案后处理过程的主要任务是关闭数据库。

选案指标的处理过程

·缺省的指标结点前期处理过程的作用是在该指标计算打分之前,检查这个指标所用到的选案变量值是否有效,这样避免产生错误的指标分数。

·缺省的指标结点后期处理过程的作用是在这个指标结点的分数计算过程处理完成以后将这个指标的有关信息、结果和解释写入相关的指标输出结果表中。

·总分的计算公式是:

总分 = 求和(指标分数 * 相应指标的权值) / 求和(所有指标的权值)。

(2)指标的“分数计算”处理过程。由于指标的分数计算过程是灵活多变的,在指标结点的“分数计算”处理过程中,系统不为它生成缺省的脚本。用户需要自己编写适当的脚本。这个处理过程的脚本一般包括这些内容:

·为了编写脚本的方便,用户自定义一些局部变量,变量名称可以是中文。

·取出需要的选案变量放在局部变量中。可用 bl() 函数取出需要的选案变量的值。

如果有,则根据计算公式进行计算。

·如果有必要,可加入相应的规则进行分数调整,最后用 fs() 函数输出分数,以便上级结点使用。

这部分一般由分支语句组成。根据不同的指标,这一部分可能变化较大。

(3)复杂的选案指标的建立方法。通过选案方法维

护工具,并运用强大的 ICL 语言,建立复杂的选案指标是不困难的。在指标结点的“分数计算”处理过程中,你可以加入能够满足各种特定需要的计算公式和打分规则。除此以外,还可以在一个指标结点中生成一些子指标结点,把一个指标分解称为几部分,然后再组织在一起。注意这个时候可能需要对相关指标结点的缺省“后期处理”过程作较大的修改。

四、小结

本文主要探讨了如何应用面向对象的知识处理系统来构造税收业务中的计算机选案专家系统。该专家系统具有如下特点:

- 用专家系统工具构造而成,技术成熟,高效而可靠。
- 采用面向对象的方式来描述问题,用户易于理解。
- 选案方法维护采用流行的可视化风格,界面友好,

使用方便,所见即所得。

- 预先定义了强大的缺省功能,用户需要做的维护工作很少;

- 处理过程用 ICL 语言及其函数库实现,可以灵活定制以适用各种特定的和复杂的情况。系统在实际运行中,得到了广东地税用户的认可和好评,现正在进一步向各级税务部门推广应用。

参考文献

- [1] 史忠植,知识工程,清华大学出版社,1988
- [2] 陆汝钤,专家系统开发环境,科学出版社,1994
- [3] 史忠植,高级人工智能,科学出版社,1997
- [4] 吴泉源,刘江宁,人工智能与专家系统,国防科技大学出版社,1995

(来稿时间:1999年2月)