数据挖掘技术在税收预测分析中的应用

Application of Data Mining Techniques in Tax Prediction Analysis

李继嵬　（清华大学软件学院 100084）
刘书明　（国家税务总局信息中心 100000）
李春平　（清华大学软件学院 100084）

摘要：本文从数据挖掘预测功能的角度介绍了数据挖掘技术，并阐述了各种方法在税务中的应用。重点介绍了两种常用于税收预测的算法，并结合税务部门的特点，给出了一个对实际工作有指导意义的预测方法。

关键词：数据挖掘　税收预测

1　引言

税收收入预测是指在一定的经济理论指导下，根据经济和税收统计资料，在定性分析基础上，运用定量方法，对未来税收收入总量和结构等发展趋势所做出的分析、判断和推测。税收预测对税收工作有着重要的作用，直接关系到决策的质量。

现在利用数据挖掘技术对税收收入进行预测已经成为必然的趋势，为了较好的解决以上问题，应在税收分析过程中充分利用数据挖掘技术，这样不但可以促进税收分析工作水平的提高，也可以给管理者提供决策依据，从而带动税收工作整体水平的提升。本文主要介绍了数据挖掘中的一些预测方法，并结合预测方法给出了在税务领域的一些具体应用。

2　数据挖掘技术

2.1　数据挖掘概述

数据挖掘通常又称数据库中知识发现（Knowledge Discovery in Databases，KDD），是自动的或方便的模式提取，这些模式代表隐藏在大型数据库、数据仓库或其他大量信息存储中的知识。数据挖掘是一门新兴的多学科交叉领域，源于数据库系统、数据仓库、统计学、机器学习、算法设计、数据可视化、信息检索和高性能计算。其他有贡献的领域有人工神经网络、模式识别、空间数据分析、图像数据库、信号处理和许多应用领域，还包括商务、经济学和生物信息学[1]。

数据挖掘涉及使用各种各样的算法来完成不同的任务。所有这些算法都试图为数据建立合适的模型。利用算法来分析数据，并确定与所分析数据的特征最符合的模型。

数据挖掘模型在本质上可分为预测型与描述型模型两类。如图1所示，每类模型下都包含一些需要用到该类模型的最常用的数据挖掘任务。

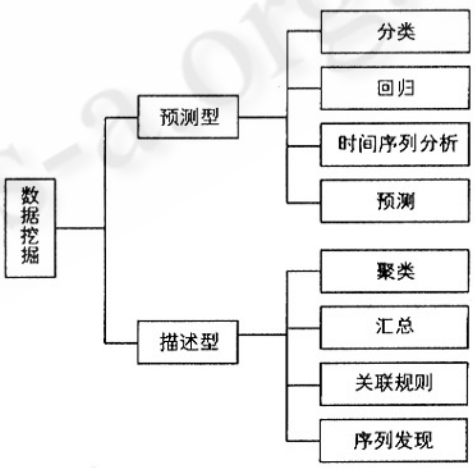

图 1　数据挖掘模型与任务

利用从不同数据中发现的已知结果，预测型模型对数据的值进行预测。预测型建模可能是基于使用其他的历史数据。例如，一户企业被归入税收欺诈队列，可能不是因为该企业自己的历史纳税数据

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

信息，而是因为其纳税额增减周期与其他有欺诈行为的企业相似，而被归入了欺诈队列。预测模型能够完成的挖掘任务包括分类、回归、时间序列分析和预测[2]。

描述型模型对数据中的模式或关系进行辨识，与预测型模型不同，描述型模型提供了一种探索被分析数据性质的方法，而不是预测新的性质。聚类、汇总（也叫特征化或泛化）、关联规则和序列发现在本质上都通常被视为是描述型的。

2.2 预测型数据挖掘任务

2.2.1 分类

分类是指将数据映射到预先定义好的群组或类。因为在分析测试数据之前，类别就已经被确定了，所以分类通常被称为有指导学习。分类是用来预测数据对象的类别，即利用现有信息预测未知事件。如预测纳税人是否有逃漏税行为，或者逃漏税等级等。

分类方法在税务稽查中是非常有用的，一户企业如何才能判定它是诚实纳税还是偷逃税的呢；上个月如实纳税不等于这个月一定没有偷逃税；企业的大小规模与行业性质在偷逃税的倾向上也没有必然的联系；这些都告诉我们有些问题没有固定的决策规则和模型可依，以往更多的是依靠稽查人员的经验及举报材料。现在可以应用分类方法在企业每月上缴的财务报表的基础上，通过对这些数据的挖掘与知识发现，建立一个纳税诚实申报判别模型，运用定性和定量相结合的方法，根据纳税人的申报资料自动判别其是否存在偷逃税行为，起到对稽查人员辅助决策的作用，可以大大增加税收征管工作的科学性，提高税务稽查的效率和准确度。

2.2.2 回归

回归是指将数据项映射到一个实值预测变量。事实上，回归涉及学习一个可以完成该映射的函数（例如线性函数、Logistic 函数等）可以拟合目标数据，然后利用某种误差分析确定一个与目标数据拟合程度最好的函数。

如某一税务部门计划在 2010 年使征收税额增长到一定的额度（当然计划应有合理性），就需要基于当年的税额和过去几年的税额，定期地预测在 2010 年时可能达到的税额。首先需要确定拟合过去历史数据的线性函数，然后利用该线性函数来预测未来的税额。根据这些数值，税务部门可以调整征收力度及稽查范围，保证计划的实现。

2.2.3 时间序列分析

在时间序列中，数据的属性值是随着时间不断变化的。一般情况下，在相等的时间间隔内（例如日、月、年等）可以得到这些数据。时间序列分析有三个基本功能。第一，使用距离度量来确定不同时间序列的相似性；第二，检验时间序列图的结构来确定（有时只是辨别）时间序列的行为；第三，利用历史时间序列图来预测数据的未来数值。

时间序列一般由两个基本要素构成：一是反映经济现象的变量的所属期，二是反映经济现象的指标及其指标值。

时间序列分析对于税收分析具有重要的作用，是常用的分析工具之一。其作用具体表现为：

- 准确描述经济现象的发展状态和结果；
- 研究现象的发展趋势和发展速度；
- 探寻经济现象的发展规律性；
- 利用时间序列所表现出来的趋势，对经济现象进行预测，是税收预测的有效方法。

2.2.4 预测

许多实际的数据挖掘应用需要基于过去和当前数据对未来数据状态进行预测。预测可以看作是一种分类。（预测任务是一种预测模型，但作为一种任务的预测不同与预测模型。）差别在于预测主要是预测未来数据的状态而不是当前状态。通常认为，当被预测的值是连续值时，称之为预测，当被预测的值是离散值时，称之为分类。预测除了可以使用时间序列分析和回归分析对未来值进行预测外，还可以使用其它技术。

对税收收入进行预测是一件很困难的事情，主要原因是影响税收收入的因素太多，但也不是一点办法也没有，我们可以通过预测技术给出一个近似值。其中一种可行的方法是在不同时期收集大量与税收相关的数据，除税务数据外还要收集下列对税收收入有影响的数据，如 GDP、消费、投资、价格、净出口等经济指标数据，结合各种数据挖掘方法，就可以根据以前的各种数据进行预测。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

3 税收预测的具体方法

3.1 常用的税收预测方法

(1) 一元线性回归预测法。一元线性回归预测是用一元线性回归模型,对具有线性趋势的税收问题,只使用一个影响因素所作的预测。比如通过税收收入与GDP的关系,建立关于税收收入与GDP的一元回归模型,用未来GDP的预测值或计划数值预测税收收入的规模。

(2) 多元线性回归预测法。多元线性回归预测法是用多元线性回归模型,对具有线性趋势的税收问题所作的预测。一元线性回归预测法和多元线性回归预测法都是线性趋势预测法,该法只适用于具有线性趋势的现象之间的关系。

(3) 非线性预测法。非线性预测法是对利用非线性模型进行预测的一系列方法的总称。最常用的非线性预测法有二次曲线预测法、指数曲线预测法等。二次曲线预测法是在确认税收与某个经济变量之间存在二次曲线趋势时,利用二次曲线模型预测税收收入的方法。指数曲线预测法是用指数曲线模型对呈固定速度增长的税收问题预测的模型。

(4) 指数平滑预测法。指数平滑是画拟合曲线的一种方法,同时还可以对将来进行预测。指数平滑就是将最近的观察数据赋予较高的权重,较早的数据赋予相对较低的权重,权重以一个常数的比率进行几何递减,使得较近的数据对将来的预测分析起的作用大一些。根据用户选择的参数不同,可以分为平稳时间序列指数平滑、趋势时间序列指数平滑,和季节周期性指数平滑。

(5) 神经网络预测法。神经网络近来越来越受到人们的关注,因为它为解决较大复杂度问题提供了一种相对来说比较有效的简单方法。神经网络可以很容易的解决具有上百个参数的问题(当然实际生物体中存在的神经网络要比我们这里所说的程序模拟的神经网络要复杂的多)。

(6) 税收预测中的具体应用。上述列出的预测方法在本质上也可以分为两大类:一类是解释性预测方法,即找出被预测量的各影响因素,建立回归分析模型;另一类为时间序列分析方法,只依赖于被预测量的历史观测数据,通过序列分析,找出其顺序变化规律。

在税收收入预测中采取的方法可以根据税收收入和其它经济因素之间的关系,用税收历史数据和各种经济指标数据,建立税收收入与GDP、工业增加值、商业增加值、消费、投资、价格、净出口等相关经济指标的多元回归模型、非线性回归模型、神经元网络或其它模型;在建模过程中要不断调整对因变量的选择,以获得一个比较好的模型。最后根据已知的数据来预测未来指定时间内的税收收入的可能值及其变化趋势。实际上为了得到满足需要的结果,我们经常采取几种方法的组合进行处理,回归与神经元网络的组合就是一种很好的选择[8]。

3.2 指数平滑的预测实例

根据平滑次数不同,指数平滑法分为:一次指数平滑法、二次指数平滑法和三次指数平滑法等。它们的基本思想都是:预测值是以前观测值的加权和,且对不同的数据给予不同的权,新数据给较大的权,旧数据给较小的权[9]。下面以某税务局的十年历史数据,应用一次指数平滑及二次指数平滑分别进行预测。

3.2.1 一次指数平滑法

设时间序列为$y_1, y_2, \cdots y_t, \cdots$,则一次指数平滑公式为:

$$S_t^{(1)} = \alpha y_t + (1-\alpha) S_{t-1}^{(1)} \quad (式4--1)$$

式中$S_t^{(1)}$为第t周期的一次指数平滑值;α为加权系数,$0<\alpha<1$

为了弄清指数平滑的实质,将上述公式依次展开,可得:

$$S_t^{(1)} = \alpha\sum_{i=0}^{t-1}(1-\alpha)^i y_{t-i} + (1-\alpha)^t S_0^{(1)} \quad (式4--2)$$

由于$0<\alpha<1$,当$t\to\infty$时,$(1-\alpha)^t\to 0$,于是上述公式变为:

$$S_t^{(1)} = \alpha\sum_{i=0}^{\infty}(1-\alpha)^i y_{t-i} \quad (式4--3)$$

由此可见$S_t^{(1)}$实际上是$y_t, y_{t-1}, \cdots y_{t-i}, \cdots$的加权平均。加权系数分别为$\alpha, \alpha(1-\alpha), \alpha(1-\alpha)^2, \cdots$,是按几何级数衰减的,愈近的数据,权数愈大,愈远的数据,权数愈小,且权数之和等于1,即$\alpha\sum_{i=0}^{\infty}(1-\alpha)^i = 1$。因为加权系数符合指数规律,且又具有平滑数据的功能,所以称为指数平滑。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

用上述平滑值进行预测，就是一次指数平滑法。其预测模型为：

$$\hat{y}_{t+1} = S_t^{(1)} = \alpha y_t + (1-\alpha)\hat{y}_t \quad (式 4--4)$$

即以第 t 周期的一次指数平滑值作为第 t+1 期的预测值。

表 1　一次指数平滑预测表　　单位：亿元

年份	序号	某局收入 Y_t	α=0.2 的预测值 $\hat{y}_t$	α=0.5 的预测值 $\hat{y}_t$	α=0.8 的预测值 $\hat{y}_t$
1995	1	18.92	21.72	21.72	21.72
1996	2	24.51	22.274	23.1125	23.951
1997	3	27.07	23.2332	25.09125	26.4462
1998	4	30.69	24.72456	27.890625	29.84124
1999	5	33.25	26.429648	30.5703125	32.568248
2000	6	35.78	28.2997184	33.17515625	35.1376496
2001	7	40.49	30.73777472	36.83257813	39.41952992
2002	8	50.31	34.65221978	43.57128906	48.13190598
2003	9	62.20	40.16177582	52.88564453	59.3863812
2004	10	70.08	46.14542066	61.48282227	67.94127624

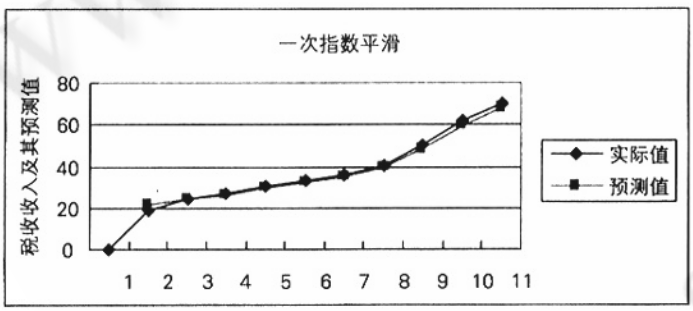

图 2　一次指数平滑散列图

$$S_0^{(1)} = \frac{y_1 + y_2}{2} = \frac{18.92 + 24.51}{2} = 21.715$$

从表中可以看出，α=0.2，0.5，0.8 时，预测值是很不相同的。究竟 α 取何值好，可通过计算它们的均方差 MSE，选取使 MSE 较小的那个 α 值。

当 α=0.2 时，$MSE = \frac{1}{10}\sum_{t=1}^{10}(Y_t - \hat{Y}_t)^2 = 156.44$

当 α=0.5 时，$MSE = \frac{1}{10}\sum_{t=1}^{10}(Y_t - \hat{Y}_t)^2 = 25.5$

当 α=0.8 时，$MSE = \frac{1}{10}\sum_{t=1}^{10}(Y_t - \hat{Y}_t)^2 = 2.85$

计算结果表明，α=0.8 时，MSE 较小，故选取 α=0.8，预测 2005 年某局税税收收入为：

$$\hat{y}_{t+1} = S_t^{(1)} = \alpha y_t + (1-\alpha)\hat{y}_t = 0.8 \times 70.08 + (1-0.8) \times 67.94 = 69.65 \text{（亿元）}$$

3.2.2　二次指数平滑法

当时间序列没有明显的趋势变动时，使用第 t 周期一次指数平滑就能直接预测第 t+1 期之值。但当时间序列的变动出现直线趋势时，用一次指数平滑法来预测仍存在着明显的滞后偏差。因此，需要进行修正。修正的方法是在一次指数平滑的基础上再作二次指数平滑，利用滞后偏差的规律找出曲线的发展方向和发展趋势，然后建立直线趋势预测模型。故称为二次指数平滑法。

设一次指数平滑为 $S_t^{(1)}$，则二次指数平滑 $S_t^{(2)}$ 的计算公式为：

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha) S_{t-1}^{(2)} \quad (式 4--5)$$

若时间序列 $y_1, y_2, \cdots y_t \cdots$ 从某时期开始具有直线趋势，且认为未来时期亦按此直线趋势变化，则与趋势移动平均类似，可用如下的直线趋势模型来预测。

式中 t 为当前时期数；T 为由当前时期数 t 到预测期的时期数；$\hat{y}_{t+T}$ 为第 t+T 期的预测值；a_t 为截距，b_t 为斜率，其计算公式为：

$$a_t = 2S_t^{(1)} - S_t^{(2)} \quad (式 4--6)$$

$$b_t = \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)}) \quad (式 4--7)$$

一般来说，若数据波动较大，α 值应取大一些，可以增加近期数据对预测结果的影响。若数据波动平稳，α 值应取小一些，使实际观察期的所有数据对预测结果的影响趋于均衡。α 值通常在 0.01—0.09 范围内选取。

在以往的工作中都是根据经验选 α 的值，即不准确又没有科学性，为此编制了一个 C 语言程序，使其能够计算出当 α=0.01，α=0.02，...，α=0.99 时，每一个 α 的平均绝对误差，得到最小误差的 α 为最终用于预测的平滑系数，就可以提高预测的准确性与可靠性。

3.2.3　实验结果分析

根据某局 2005 年的实际税收数 77.57 亿元，对预测结果分析如下：

（下转第 68 页）

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

(上接第 64 页)

表 2　两种方法对比分析表　　单位:亿元

方法	一次指数平滑	二次指数平滑
实际数	77.57	77.57
预测数	69.65	78.04
方差	62.7264	0.2209
绝对误差	7.92	-0.47
相对误差	0.102101328	-0.006059043

从表 2 可以看出二次指数平滑预测法的绝对误差小于 0.5,相对误差小于 0.01,能够满足税收计划及决策支持工作的实际需要。

4　结论

越是精密的技术越是脆弱,正如导弹的杀伤力远远大于长矛大刀,但导弹出现错误和无法使用的可能也远远大于长矛大刀。税务部门应用数据挖掘技术也是一样,为了避免或减少错误的发生,要求实施项目的核心人员必须首先是熟悉数据,其次熟悉业务,其熟悉业务的程度可以说和挖掘的成效成正比。

采用数据挖掘技术作为税收收入定量预测的方法,目前还不是很完善与成熟,需要我们不断的探索,实践,探索,再实践直到达到理想的要求。

参考文献

1 Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques. 2000:185 ~ 217。

2 Margaret H. Dunham 著,郭崇慧、田凤占、靳晓明等译,数据挖掘教程,清华大学出版社,2005 - 5:4 ~ 8。

3 Olivia Parr Rud 著,朱扬勇、左子叶、张忠平等译,数据挖掘实践,机械工业出版社,2003 - 9:10 ~ 15。

4 吴维扬,经济预测及案例分析,中国信息大学文库,1995 - 5:32 - 48。

5 徐国祥,统计预测和决策,超星数字图书馆,1994 - 4:91 - 109。

6 沈永淦、周格非,实用经济预测,1986:115 ~ 168。

7 国税总局信息中心, 国家税务总局税收宏观决策支持系统概要设计说明书, 2005 - 2:35 ~ 47。

8 崔德光、吴淑宁、徐冰,空中交通流量预测的人工神经网络和回归组合方法,清华大学学报,2005 - 1:96 ~ 99。

9 卢小广,统计学教程,清华大学出版社,2006 - 1:267 ~ 270。

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn