神经网络与领域知识结合的纳税评估预警模型

徐 戎[1]，王文杰[1]，周四新[2]

(1. 中国科学院研究生院 北京 石景山区 100039; 2. 中国人民大学财政金融学院 北京 海淀区 100872)

【摘要】纳税评估是一项重要而复杂的工作。针对目前尚无十分有效的纳税评估预警模型的情况，提出了将神经网络和领域知识结合建立纳税评估预警模型的方法，利用基于神经网络的方法选出有涉税疑点的企业，在领域知识的指导下结合统计分析方法，解决了预警模型无疑点指向性的问题。通过建立行业的纳税评估预警模型，并进行验证分析，表明该方法是可行的。

关 键 词 神经网络； 预警模型； 统计分析； 纳税评估
中图分类号 TP183; F813 **文献标识码** A

Risk Warning Model of Tax Payment Evaluation Based on Neural Network and Domain Knowledge

XU Rong[1], WANG Wen-jie[1], and ZHOU Si-xin[2]

(1. Graduate University of Chinese Academy of Sciences Shijingshan Beijing 100039;
2. School of Finance, Renmin University of China Haidian Beijing 100872)

Abstract A risk warning model of tax payment evaluation based on neural network and domain knowledge is presented in this paper. The enterprises that have the most possibility of tax dodging can be found by applying neural network methods. The doubtful points in tax payment of these enterprises are evaluated by statistical analysis methods. A case study shows the feasibility of the model.

Key words neural network; risk warning model; statistical analysis; tax payment evaluation

纳税评估是一项国际通用的税收管理制度，建立纳税评估预警模型主要是为了能有效地选取评估对象，对纳税评估工作提供指导作用。我国对纳税评估预警模型的研究还处于起步阶段，比较普遍的方法是利用税务人员的经验建立本地区的纳税评估指标体系进行预警。随着信息技术在税收领域的运用，目前已有一些将计量经济方法和智能信息技术应用于纳税评估领域的尝试，如将Logit回归分析方法引入纳税评估、基于C4.5挖掘算法对行业理论税负进行测算[1]、利用遗传算法优化BP神经网络的方法对纳税人是否诚信的判断[2]、利用聚类分析方法找出纳税异常企业[3]等。针对纳税评估工作的复杂性，如何有效地选取涉税疑点大的评估对象、也为评估人员的实际工作提供涉税疑点的指向性，目前仍处于探索阶段，尚无很理想的解决办法。本文提出了将神经网络与领域知识结合建立纳税评估预警模型的方法，即通过BP神经网络模型筛选出评估对象，根据领域知识利用统计分析方法对选取的评估对象进行分析，弥补了神经网络输出不可解释性、预警模型无疑点指向性的缺点，为分行业纳税评估预警模型的建立提供了一种可行的途径。

1 神经网络与领域知识结合的纳税评估预警模型

1.1 可行性分析

税收领域涉及的因素多、关系复杂，利用传统方法建立的纳税评估预警模型存在着难以处理高度非线性的复杂信息、缺少自适应能力等问题。人工神经网络利用样本集进行学习，并通过分布式并行计算解决问题，具有自学习、自适应能力，适合处理高度复杂和非线性的问题，因此，利用神经网络模型进行纳税评估预警比传统方法有许多优势。纳税评估与企业财务风险预警有很多类似之处，国外应用神经网络对企业财务风险预警的实证结果表明，人工神经网络模型在预测精度上优于多元线性回归模型、Logit回归模型等[4]。神经网络具有高精

收稿日期：2008－09－15；修回日期：2008－11－27
作者简介：徐 戎(1969－)，女，硕士生，主要从事智能信息处理方面的研究.

度的优势，但也有不透明性和难解释性的问题，对选出的评估对象难以提供涉税疑点的指向性。

税务部门在多年的稽查中积累了规律性的知识，既含有微观经济学的因素又是经验的总结。根据领域知识，利用统计分析方法对选取的评估对象进行分析，可以提供对涉税疑点的指向性。

1.2　BP神经网络及网络设计

BP(backpropagation)网络是目前应用最广泛的神经网络系统，由输入层、隐层和输出层组成，各层由若干个节点构成。BP 学习算法由正向传播和误差反向传播两个过程组成，在正向传播过程中，输入信息从输入层经隐含层传到输出层，如果与期望值一致，学习算法结束；如果有误差，则误差反向传播，沿原先的连接通路返回，逐层修改各层神经元的权值和阈值，使网络对输入信息经过计算后所得到的输出，达到期望的误差要求。

神经网络的学习问题可表述为：根据 n 个独立同分布的观测样本$(x_1, y_1), (x_2, y_2),\cdots,(x_n, y_n)$在一组函数$\{f(x,w)\}$中求最优的函数 $f(x,w_0)$，使得预测期望风险[5]最小，即：

$$R(w)=\int L(y, f(x,w))\mathrm{d}p(x,y) \tag{1}$$

式中　$\{L(y, f(x,w))\}$为给定输入 x 时，目标值 y 与网络输出 $f(x,w)$之间的差异。

用已知的训练样本定义经验风险[5]为：

$$R_{\mathrm{emp}}=\frac{1}{N}\sum_{i=1}^{N}L(y_i, f(x_i,w)) \tag{2}$$

用对参数 w 求经验风险 $R_{\mathrm{emp}}(w)$的最小值代替求期望风险 $R(w)$的最小值即为经验风险最小化(empirical risk minimization，ERM)原则。

当样本数目 l 趋于无穷时，经验风险 $R_{\mathrm{emp}}(\alpha)$一致收敛于真实风险 $R(\alpha)$。但训练样本的数目往往是有限的，由此导致经验风险与期望风险的不一致。尽管训练的网络具有最小的 ERM 值，但推广能力却不够好，这就是“过适应”问题[6]。

在有限样本的情况下，需要通过合理设置网络结构达到学习精度和推广性的兼顾[7]，因此，应控制输入神经元的维数和隐层数，降低模型的复杂度。

有关研究表明，一个两层网络只要在其隐层有足够的神经元，并且隐层神经元的传输函数是S型类型，便可以逼近任何实际的函数[8]。

1.3　统计分析方法

统计分析方法是预测、预警领域的经典数学方法。描述性统计通过考察某个统计量的频数、集中趋势、离散程度等属性来发现样本中存在的异常情况。回归分析方法是使用原始数据对目标函数进行拟合，揭示因变量与自变量间的关系。通过已知的样本信息得到样本回归函数，估计出总体回归函数，进而对异常值进行预警。

1.4　建立预警模型

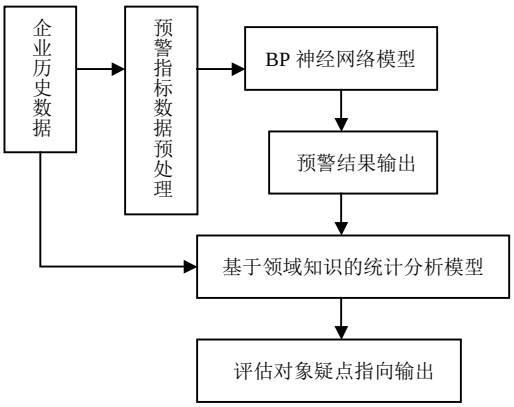

图 1　基于神经网络和领域知识的预警模型

不同行业的差异性很大，本文认为有效的纳税评估预警模型应该分行业建立。本文建立的模型如图1所示，分为以下步骤：

(1) 选取预警指标并进行预处理。

预警指标的选取应以比率类指标为主。在获得的训练样本数量较少时，选取指标的个数不宜过多，并且应包含尽可能多的有效信息。考虑到指标的预警值不易设定，而模型是在同地区同行业企业中选取评估对象，可以将指标值与行业均值进行比较，对指标值进行预处理。

(2) 基于 BP 神经网络的预警模型。

构建有一个隐含层的 BP 神经网络预警模型，建立预测函数 $\boldsymbol{Y}=f(\boldsymbol{X})$，其中，$\boldsymbol{X}(\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\})$是模型的输入，即对企业预警指标值处理后的数据；$\boldsymbol{Y}$是模型的输出，即对企业疑点确认度的判定。将税务稽查部门对企业的稽查数据作为训练和测试样本，考虑到实际稽查结果可能存在一定的偏差，对样本进行主动选取，通过对模型的训练建立 BP 神经网络预警模型。利用建立的 BP 神经网络模型，可以实现对企业按疑点确认度的分类。

(3) 统计分析模型提供评估对象的疑点指向性。

对神经网络模型的输出结果进行后处理，基于领域知识选取适于确定该行业纳税疑点指向的指标，利用统计分析方法提供涉税疑点指向性。

在某一行业同类企业数量足够大的情况下，可以认为选取指标的值趋于正态分布，根据该指标的同类企业行业均值μ求出其置信度为σ□的置信区

间，通过评估对象指标值的偏离情况提供疑点指向性。

2 实 例

本文以酒店住宿业为例建立该行业的纳税评估预警模型，数据样本来源于北京市某区的酒店住宿企业。

2.1 预警指标选取

模型选取了主营业务利润率、利润率、成本费用利润率、净资产收益率、税负率、营业税税负率、流动资产周转率共 7 个比率类指标，计算样本的指标值与行业均值的差异值作为 BP 模型的输入变量。

2.2 BP 模型构建

2.2.1 网络结构设计

输入层神经元节点数为7，输出层神经元节点数为3，实现对疑点确认度的三类模式分类，定义输出层对样本的输出向量为$\boldsymbol{y}=[y_1\ y_2\ y_3]$，输出值$\boldsymbol{y}$=[l 0 0]代表企业存在涉税问题可能性较大，$\boldsymbol{y}$=[0 1 0]代表企业存在涉税问题可能性一般，$\boldsymbol{y}$=[0 0 1]代表企业存在涉税问题可能性很小。

隐含层节点数的确定，参考了以下公式：

$$n=\sqrt{n_i+n_0}+a \tag{3}$$

式中 n为隐层节点数；n_0为输入节点数；n_i为输出节点数；a为1～10间的常数[9]。

通过反复试验，确定了隐含层的节点数为5个，最终建立的BP网络模型结构是7×5×3。模型的表达式为：

$$y_n=f\left(\sum_{j=1}^{5}v_j f\left(\sum_{i=1}^{7}w_{ji}x_i+b_j\right)+b_n\right)\quad n=1,2,3 \tag{4}$$

传递函数$f(n)$均采用Logistic函数，即：

$$\mathrm{logsig}(n)=\frac{1}{1+\exp(-n)} \tag{5}$$

2.2.2 样本训练及结果检验

本文将接受过税务稽查的企业被检查年度的数据作为训练和测试样本，把对企业的稽查结论转化为相应的输出向量作为输出的目标值，参考领域专家的意见选取了结论可信度较大的60个样本作为训练和测试样本。

在Matlab7.0环境下选取Levenberg-Marquardt学习方法进行训练。选取误差精度为$e=10^{-3}$，性能函数为mse，$\mu=0.01$，经过257步迭代后，网络达到稳定状态并收敛，如图2所示。

为了检验实际输出对期望输出的拟合程度，本文对训练样本进行仿真。仿真结果表明，预警模型对训练样本集的分类准确度达到96%。

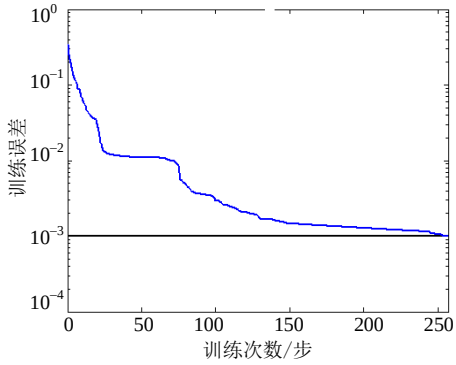

图 2 L_M 训练

将作为测试样本的数据代入同一个模型，利用仿真程序对测试样本进行分类，结果如表1所示。

表1 稽查结果与样本输出对比

稽查结果	模型输出	预警结果	对比
(0,0,1)	(0.000 1,0.000 0,0.998 1)	可能性小	
(0,1,0)	(0.000 1,0.993 6,0.000 0)	可能性一般	
(0,1,0)	(1.000 0,0.000 0,0.000 0)	可能性大	误判
(0,0,1)	(0.982 6,0.161 5,0.000 0)	可能性大	误判
(0,0,1)	(0.068 8,0.000 0,0.999 8)	可能性小	
(1,0,0)	(1.000 0,0.000 0,0.008 2)	可能性大	
(1,0,0)	(1.000 0,0.000 0,0.000 0)	可能性大	
(1,0,0)	(1.000 0,0.000 0,0.000 6)	可能性大	
(1,0,0)	(1.000 0,0.000 1,0.000 7)	可能性大	
(1,0,0)	(1.000 0,0.000 0,0.013 9)	可能性大	
(0,0,1)	(0.000 1,0.000 0,0.995 9)	可能性小	
(0,1,0)	(0.050 9,1.000 0,0.000 0)	可能性一般	

由表1可知，大部分分类正确，没有将有涉税问题的企业误判为无涉税问题企业的情况；两个误判是将稽查结果未发现涉税问题和涉税问题较小的企业误判为涉税问题较大的企业。考虑到在税收稽查工作中存在纳税人有涉税问题，但未能通过税务稽查发现的情况，因此产生的误判也与实际情况比较相符。纳税评估预警模型建立的目的是为了筛选可能存在涉税问题的纳税人，从这个意义上讲，本文所建立模型的预警效果应该是比较好的。

2.3 基于领域知识利用统计方法对评估对象进行分析

经均值检验对企业分类后，本文选取了利润率、流动资产周转率、主营业务费用率等多个指标进行分析。以利润率为例，经过 SAS 软件检验认为利润率趋于正态分布，均值为−0.076 7，标准差为 0.317 8。

经检验作为训练和测试样本需要补缴企业所得税的的企业中，有 70%利润率在置信度为 68%的置信区间的下方；如果企业同时存在隐瞒营业收入的问题，则其利润率有可能不在置信区间下方。利用流动资产周转率、主营业务费用率、货币资金变化率与营业收入变化率配比等指标，可以帮助发现企业是否有未结转当期收入、多列成本费用、隐瞒营业收入的问题。

本文对主营业务收入与费用进行回归分析，得到模型显著性检验的 F 值为 404.29，拒绝模型非显著的概率小于 0.000 1，回归模型是显著的，主营业务收入与费用的相关系数为 0.759 5，拟合的效果较好，系数的 t 检验均拒绝等于零的假设。利用主营业务收入与费用的回归模型，可以帮助提供评估对象多列成本费用，隐瞒营业收入等问题的指向性。因此，基于领域知识利用统计分析方法对筛选出的评估对象的情况进行分析，可以较好地解决疑点指向性的问题。

3　结论和展望

将神经网络与领域知识结合建立分行业纳税评估预警模型的方法能较好地对有涉税问题的企业产生预警，并为实际评估工作提供一定的疑点指向性，为解决当前纳税评估工作中选户和涉税疑点指向的问题，提供了一种可行途径。

网络结构的设计、训练样本的数量和质量对模型的泛化能力都十分重要，即：(1) 需要合理地设计网络结构。(2) 可以将准确度较高的税务稽查和评估案例建立训练样本库，也可以考虑将先验知识表示成虚拟样本，再直接用于网络训练[10]。神经网络集成通过训练多个神经网络，能提高神经网络系统的稳定性和泛化能力[11]，可以考虑在模型中运用神经网络集成的技术。

由于涉税问题的判定不仅仅是单纯的税收财务指标，还有复杂的人为、环境因素，模型是以特定样本指标反映总体实际情况并进行预测，因此难免有一定的误差，可以再利用人为因素对结果予以辅助修正。

参　考　文　献

[1] 倪　涛, 刘　耀. 基于C4.5挖掘算法的纳税评估模型设计[J]. 现代计算机, 2007, (9): 83-86.
NI Tao, LIU Yao. Design of tax payment evaluation model based on C4.5 mining algorithm[J]. Modern Computer, 2007, (9): 83-86.

[2] 蔡伟鸿, 郭陈熹. 遗传算法优化BP神经网络在纳税评估中的应用[J]. 汕头大学学报(自然科学版), 2008, 23(2): 62-63.
CAI Wei-hong, GUO Chen-xi. Application of tax assessment based on genetic algorithm optimized BP neural network[J]. Journal of Shantou University(Natural Science Edition), 2008, 23(2): 62-63.

[3] 张光前, 王久铱, 周宽久. 基于领域知识的纳税评估方法研究[J]. 数理统计与管理, 2007, 27(2): 286-287.
ZHANG Guang-qian, WANG Jiu-yi, ZHOU Kuan-jiu. Study on method of tax checking based on domain knowledge[J]. Application of Statistics and Management, 2007, 27(2): 286-287.

[4] HUANG Chin-shen, DORSEY R E, BOOSE M A. Life insurer financial distress prediction: a neural network mode [J]. Journal of Insurance Regulation, 1994,13(2): 131-167.

[5] VAPNIK V N. Principle of risk minimization for learing theory[J]. Advances in Neural Information Processing Systems, 1992, 4: 831-838.

[6] 黄　华, 罗四维, 刘蕴辉, 等. 人工神经网络知识增殖性分析[J]. 计算机研究与发展, 2005, 42(2): 224-229.
HUANG Hua, LUO Si-wei, LIU Yun-hui, et al. Knowledge increase ability analysis on artificial neural network[J]. Journal of Computer Research and Development, 2005, 42(2): 224-229.

[7] 阎平凡, 张长水. 人工神经网络与模拟进化计算[M]. 第2版. 北京: 清华大学出版社, 2005.
YAN Ping-fan, ZHANG Chang-shui. Atrificial neural networks and evolutionary computing[M]. 2nd. Beijing: Tsinghua University Press, 2005.

[8] MARTIN T, HAGAN H B, BEALE D M H. Neural network design[M]. Translated by DAI kui. Beijing: China Machine Press, 2006.

[9] 周开利, 康耀红. 神经网络模型及其MATLAB仿真程序设计[M]. 北京: 清华大学出版社, 2005: 89-90.
ZHOU Kai-li, KANG Yao-hong. Neural network model and simulation program design based on matlab[M]. Beijing: Tsinghua University Press, 2005: 89-90..

[10] NIYOGI P. Incorporating prior information in machine learning by creating virtual examples[J]. Proc IEEE, 1998, 96: 2196-2209.

[11] HANSEN L K, SALAMON P. Neural networks ensembles [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 1990, 12(10): 993-1001.

编　辑　黄　莘