基于概率神经网络和K-means算法的纳税评估

赵 雷[1],张延荣[2]

(1.河北工程大学 经济管理学院,河北 邯郸 056038;2.邯郸市第一医院,河北 邯郸 056002)

[摘 要]纳税评估工作是一项难于建立准确数学模型的复杂系统,同时又是一个典型的模式识别问题。用神经网络方法进行纳税评估有其独特的优越性。运用PNN算法很大程度地依赖于训练样本对象的选取。选取的这些样本能否反映总体的信息特征决定了分类器的识别效果。文章运用K-means算法对纳税人信息样本进行聚类,找出聚类中心点,以此为基础来选择样本作为PNN的训练样本,从而达到对PNN算法的优化。研究结果表明这种改进后的PNN算法分类效果好,对于纳税评估有其应用价值。

[关键词]纳税评估模型;概率神经网络;模式分类;K-means算法

[中图分类号] C931 [文献标识码]A [文章编号]1673-9477(2011)01-0027-02

概率神经网络(PNN)是Specht[1]于1990年提出的一种人工神经网络模型。它是基于贝叶斯决策[2]和Parzen窗概率密度函数估计方法的一种适用于模式分类的径向基神经网络。PNN算法通过把每类的后验概率的估计转化为先验概率的估计,根据事件先验概率发生的大小,进而对未知模式进行判决。一般来说,样本的概率密度函数(probability density function,PDF)通常不能准确获得,只能根据现有样本特征求其统计值。1962年,Parzen提出了一种从已知随即样本中估计概率密度函数的方法,只要样本数目足够多,该方法所获得的函数可以连续平滑地逼近原概率密度函数。从结构上看,概率神经网络属于多层前向型神经网络,通常由隐层、输出层两个神经元层组成。相对于BP神经网络而言往往需要更多的神经元,但是它的训练速度更快,在输入向量样本数目较多的情况下,PNN网络的分类效果是很好的。

本文运用PNN方法,针对文献3中的样本数据进行两类模式分类,考虑文献3中的纳税人信息指标:总资产报酬率、流动资产周转率、存货周转率、股东权益比率和固定资产成新率。运行结果表明,PNN对两类模式分类准确率达到了80.26%。进一步研究表明,运用K-means算法对训练样本进行聚类,并利用聚类中心来定义PNN隐中心矢量,一方面减少了PNN神经元个数,另一方面提高了对测试样样本分类的准确率。

一、纳税评估模型的建立

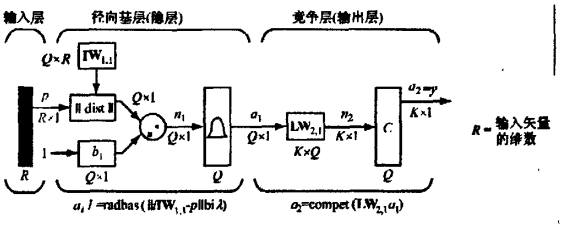
(一)纳税评估问题描述

根据国家《纳税评估管理办法(试行)》的解释,纳税评估是指税务机关运用数据信息对比分析的方法,对纳税人和扣缴义务人纳税申报情况的真实性和准确性作出定性和定量的判断,并采取进一步征管措施的管理行为。纳税评估对象的选取是纳税评估工作的起点,本文运用PNN方法建立的纳税评估模型,对税收不遵从纳税人的判断有较好的泛化能力。

本文采用文献3中68户没有违规情况的纳税企业和8户经稽查已证明发生偷逃税企业的会计信息作为样本数据.考虑样本的5个财务指标:总资产报酬率、流动资产周转率、存货周转率、股东权益比率和固定资产成新率5个指标。对于8户逃税企业,随机选取其中4户作为训练样本,其余作为测试样本。对于68户没有违规情况的企业,选择30户作为训练样本,其余作为测试样本。

(二)概率神经网络算法描述

概率神经网络是径向基神经网络的一种,通常用于模式分类的问题。在实际应用中,当概率神经网络获得一个输入时,第一层神经元计算输入向量与输入样本向量的距离,并产生一个向量,此向量的各个元素表征输入向量与输入样本向量的接近程度。第二层神经元将与输入向量的各种类别综合起来,产生一个表征概率的输出向量。最后,一个竞争型的传递函数在输出端选择具有最大概率的输入向量类别,产生输出1,而其他类别的输入向量则产生输出0。其网络结构如下图所示。



R为输入向量元素的数目;Q为输入目标样本的数目,也是第一层神经元的数目;K是输入向量类别的数目,也是第二层神经元的数目。

(三)两类模式样本的选取

对于两类模式分类问题,我们定义逃税企业为1类,没有逃税的企业为2类。定义1类错误为将没有逃税的企业误判为逃税企业,定义2类错误为将逃税的企业误判为没有逃税的企业。

二、模型的检验及结果分析

单独运用pnn对两类模式训练样本集和测试样本集分别进行判别的仿真结果如表1所示:

[收稿日期]2011-03-10

[作者简介]赵雷(1982-),男,河北磁县人,硕士研究生,研究方向:技术经济及项目管理。

表1

Pnn方法	第1类错误	第2类错误	总误判	正确率
训练样本集(34个)	0	0	0	100%
测试样本集(42个)	10	2	12	71.42%
总体(76个)			12	84.21%

用kmeans算法对训练样本中30户没有逃税的企业进行聚类,找出20个聚类中心向量,作为PNN的输入样本向量。经过反复聚类得到仿真结果如表2所示:

表2

Pnn方法	第1类错误	第2类错误	总误判	正确率
训练样本集(34个)	0	0	0	100%
测试样本集(42个)	15	0	15	64.28%
总体(76个)			15	80.26%

由表2得出运用K-means算法降低了第2类错误。

散布常数选择对于概率神经网络的创建是有很大影响的,为了测试在不同散步常数情况下的网络性能,在MATLAB中运用for循环编写如下程序:

```
for i = 1:5
net = newpnn(P,T,i/5);
temp = sim(net,P)
yc = vec2ind(temp)
end
```

利用K-means算法进行聚类确定训练样本,在此基础上对散步常数进行调试。选出最优的一组,仿真结果如表3所示:

表3

Pnn方法	第1类错误	第2类错误	总误判	正确率
训练样本集(34个)	0	0	0	100%
测试样本集(42个)	8	0	8	80.95%
总体(76个)			8	89.47%

由表3得出在运用MATLAB对散步常数进行优化后,使第一类错误降低。同时,总的误判数也降低。

三、结论与探讨

概率神经网络应用范围十分广泛,目前已被应用于电路故障分析[4],模式识别[5],水质评价[6]等方面。本文建立的基于概率神经网络的纳税评估模型具有学习过程简单,训练速度快,在输入向量样本数目较多的情况下效果好的优点。由于其网络结构只有两层,且在运算时不需要返回对网络权值进行修改,因此执行速度快。利用K-means算法对PNN的训练样本进行优化不仅减少了训练样本的数量,更提高了PNN预测精度。同时通过对散布函数的不同取值,网络预测的精度进一步提升。本文所提出的方法利用MATLAB软件实现可以做为纳税评估工作中对可能逃税企业的初步筛选的重要参考,进而提高纳税评估工作的效率。

[参考文献]

[1] Specht. D. F. Probability Neural Networks. Neural Networks. 1990,3:109~118
[2] MacKay. D. J. C. Bayesian Methods for Neural Networks:Theory and Applications. Neural Computation. 1995:448~472
[3] 王银. 粗糙集和支持向量机在纳税评估中的应用研究[D]. 重庆大学,2008
[4] 张洪波. 基于主成分的概率神经网络模拟电路故障诊断的研究[D]. 湖南大学,2008
[5] 汪满满. 结合多阶段FCM和差异算法的概率神经网络模式识别方法研究[D]. 重庆大学,2008
[6] 陈永灿,陈燕,郑敬云等. 概率神经网络水质评价模型及其对王峡近坝水域的水质评价分析[J]. 水力发电学报,2004,23(3):1-4.

[责任编辑:陶爱新]

Tax assessment based on PNN and K-means algorithm

ZHAO Lei[1], ZHANG Yan-rong[2]

(1. College of Economics and Management, Hebei University of Engineering, Handan 056038, China;
2. Handan First Hospital, Handan 056001, China)

Abstract: Tax assessment is a complex system which is hard to build a accurate mathematical model, furthermore, it is a typical recognition problem of model. to do tax assessment by using probabilistic neural network has its special advantages. the use of PNN algorithm depends greatly on the choosing of sample. whether the chosen sample can reflect the general information characteristics determines the recognition effects of classifying machine. In this article, K-means method is used to get the cluster center of tax payers' information sample to choose sample as the training sample of PNN, thus reaching optimization of PNN algorithm. The research result shows that the classifying effects of improved PNN algorithm is good and has its values of tax assessment.

Key words: model of tax assessment; probabilistic neural network; pattern classification; K-means algorithm

(上接第17页)

Study on urbanization as a driving force of industrial aggregation of Hebei province

SUN Hong-zhe, ZHANG Hong-mei

(Economics and Management School, Hebei University of Engineering, Handan 056038, China)

Abstract: Industry aggregation is an important factor in generating mass economy and creating the core of regional competitive power to which urbanization serves as a driving force. This paper studys the theory of urbanization driving industry aggregation; tests the promoting effect of urbanization, analyzes the existed problems; and promots the route of industry aggregation driven by the upspeading urbanization.

Key words: urbanization; industry aggregation; mechanism