

基于 SOM 的预警模型研究与应用^{*}

彭 岩¹ 王万森¹ 涂序彦²

(首都师范大学信息工程学院 北京100037)¹ (北京科技大学信息工程学院 北京100083)²

摘 要 在预警系统中建立有效的风险预警模型是十分重要的。我国进行风险预警起步较晚,历史数据残缺不全,预警时有些因素带有模糊性,而神经网络带有高度并行处理信息的机制且有自学习、自适应能力,因而使系统灵活性更强。本文提出并建立了一种基于 SOM 神经网络的风险预警模型,解决了模型训练和样本选用中的一些问题。最后用实际数据对模型进行了检验,对预警结果进行了分析。

关键词 预警模型,自组织映射,人工神经网络

Study on the Model of Early Warning Based on SOM

PENG Yan¹ WANG Wan-Sen¹ TU Xu-Yan²

(Information Engineering College, Capital Normal University, Beijing 100037)¹

(Information Engineering College, USTB, Beijing 100083)²

Abstract In the Risk Early-Warning System, risk prediction model is very important that, it is regarding as the core of risk early warning. The research of early warning is very late in China, and do not have complete data in history. So early warning has its uncertainty. Compared with traditional models, neural network has self-study, self-organizing as well as self-adaptive advantages. This paper introduces SOM neural network into the Risk Early-Warning System. The methods are improved in select sample data and train the network. Experimental results show that the SOM model is accurate and can give visual prediction results as well.

Keywords Early warning model, SOM, ANN

1 引言

人工神经网络(ANN, Artificial Neural Network)是对生物神经网络系统的模拟,其信息处理功能是由网络单元的输入输出特性(激活特性)、网络的拓扑结构(神经元的联接方式)所决定的。ANN 对问题的求解方式与传统方法不同,它是经过训练来解答问题的。训练一个 ANN 是把同一系列的输入例子和理想的输出作为训练的“样本”,根据一定的训练算法对网络进行足够的训练,使得 ANN 能够学会包含在“解”中的基本原理。当训练完成后可以用来求解相同的问题。作为研究复杂性的有力工具,神经网络技术近年来在模式识别与分类、识别滤波、自动控制、预测等方面已展示了其非凡的优越性,特别是能处理任意类型的数据,这是许多传统方法所无法比拟的。

这里,我们将 SOM 网络用于信贷风险预警,并对其中的迭代方法进行了改进。

2 Kohonen 自组织模型

人脑是由大量细胞组成的,在对人类的神经系统及脑的研究中,人们发现:在空间中处于不同位置的脑细胞区域控制着人体不同部位的运动。依据大脑对信号处理的特点,在1981年,芬兰的 T. Kohonen 提出了一种神经网络模型,也就是自组织特征映射模型 SOM (Self-Organizing feature Map)^[1]。他认为神经网络中邻近的各个神经元通过彼此侧向交互作用,相互竞争,自适应发展成检测不同信号的特殊检测器,这就是自组织特征映射的含义。一个神经网络接受外界的输入模式时,将分为不同区域,各区域分别对不同的输入特征敏

感,且这一过程是可以自动完成的。

Kohonen 的思想在本质上是希望解决有关外界信息在人脑中自组织地形成概念的问题。对于一个系统来说,就是要解决一个系统在受外界信息作用时在内部自组织地形成对应表示形式。这包括神经网络的权系数调整。神经网络的自调整过程和大脑的自组织过程是相仿的。由于神经网络是由可以自调整的神经元组成,因此可以自组织成对外界信息中某一种特征敏感的形式^[2]。

3 预警实例

3.1 概述

信贷是银行业的主要收益来源,同时也是风险最大的经营产品。信贷风险预警是一个复杂的过程,它通过一系列技术手段,提早发现和判别风险来源、风险范围、风险程度和风险趋势,并发出相应的风险警示信号^[3]。实际上在贷款发放时一般就将贷款预分为正常贷款和关注贷款。这种预分类的过程实际上就是贷前的信贷风险分析过程。它为贷与不贷、贷多贷少、贷款利率和贷款安排提供决策支持。银行管理者就能及时运用一系列方法来避免或减少贷款损失。显然,在风险预警系统中建立有效的风险预警模型是十分重要的,它是风险预警的核心,对商业银行进行信贷决策、制定信贷政策及及时采取规避风险的措施均有直接影响。

风险预测的显著特点是以借款人偿还贷款能力为核心^[4]。本文主要讨论有关财务状况的预警信号。因为财务因素与其盈利能力、资产运用能力、资本结构和现金流量等因素密切相关,所以本文从分析各种财务比率入手,重点关注借款人现在和将来的偿债能力。

^{*} 本文得到北京市教育委员会科技发展计划(No. KM200410028012)资助。彭 岩 副教授,博士,主要研究领域:人工智能理论及应用,计算机网络。王万森 教授,研究方向为人工智能,知识工程。涂序彦 教授,研究方向为人工智能及其应用。

3.2 样本数据

为了检验模型对信贷风险分析的有效性,特别是在预测借贷人是否具有偿付能力中的效果,我们选用129家公司的实际数据进行预测实验。

本文采用的样本总数为129家公司中的五项财务比率数据。其中,破产的公司有65家,其余是有偿付能力的公司。为了建模和检验的需要,将样本随机分为两组,其中:

第一组由74家公司组成,用于训练。其中36家为有偿付能力的公司,38家为破产的公司。

第二组由55家公司组成,用于测试模型。其中28家为有偿付能力的公司,27家为破产的公司。

4 SOM 网络模型构建

4.1 模型构建与训练

SOM 模型选用 SOM 图的二维拓扑结构,即由两层神经网络组成(如图1所示)。

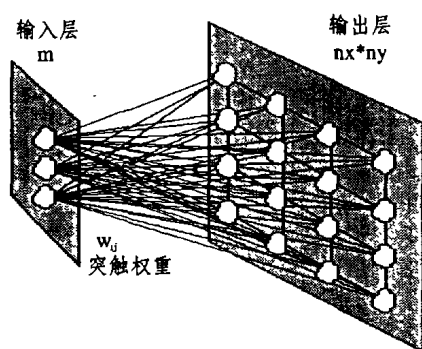


图1 二维SOM网络

输入层中的神经元数量与其所包含的变量数一致,其功能是获取数据。设 m 为输入层中是神经元的数量;同时设 $n_x * n_y$ 为神经元输出层的数量,这一层排成一个 x 行和 y 列的矩形,称为“映射”图^[5]。SOM 试图用这样的方法将多维输入空间(在我们的应用中是各种财务比率指标),投射到输出空间。以这种方式,输入模式变量中值相似地在所创建的映射中彼此接近。每个神经元学习辨识一个特定类型的输入模式。在映射图中彼此接近的神经元会辨识相似的输入模式,这些输入在所生成的图像中就会彼此接近。这样,输入空间的主要拓扑结构在输出空间中得以保存。为了达到这一目的, SOM 使用了一个竞争的算法“胜者取全部”。最初权重 w_{ij} 是给予的随机值。这些值会在运算过程(训练)中被不断修正。

训练过程是:在输入层输入财务比率,一次一个公司。设 R_{ik} 是 k 公司比率 i 的值,这项比率将会被神经元 i 读取。算法每次在输出层中取一个神经元,并且计算 Euclidean 距离做为相似性度量。

$$d(j, k) = \sqrt{\sum_{i=1}^n (r_{ik} - w_{ij})^2} \quad (1)$$

输出神经元中 $d(j, k)$ 最小的是“赢原”。设该神经元是 k^* 。算法这时以减少 $d(j, k^*)$ 距离的方式改变突触权重 (Synaptic Weight) w_{ij} 。权值的修正依赖于已经进行的迭代次数以及不同 r_{ik} 和 w_{ij} 之间的绝对值。但是其他的权重也做了相应调整,这取决于其与“赢原” k^* 的距离及已经进行了的迭代次数。

这个过程不断重复直到全部的训练停止。训练完成后,权重就固定了,网络则可以投入使用。这时,当一个新的模式输入的时候,计算输入矢量与所存储的权重矢量之间的距离,取

权重与输入矢量距离最近的神经元,映射中的激活神经元即是输出,被激活的神经元即指示出该输入矢量所属的类别。

SOM 训练算法如下:

1. 权值初始化

对连接输入神经元和输出神经元之间的权系数设定为小的随机数 a , 一般有: $0 < a < 1$, 同时,设定邻近区域的初始半径 N_c ; 时间计数置 $t=0$;

2. 在样本集中随选择一个样本 X 作为输入 $x_k = \{X_{1k}, X_{2k}, \dots, X_{nk}\}$ $k=1, 2, \dots$

3. 求模式 X_k 和所有的输出神经元的距离

for $j=1$ to p do

$$d_{jk} = \sqrt{\sum_{i=1}^n [X_{ik}(t) - W_{ij}(t)]^2}$$

4. 选择最优匹配的输出神经元 C (竞争过程)

$\min = d_{1k}; c=1;$

for $j=1$ to p do

if $d_{jk} < \min$ then $\{\min = d_{jk}; c=j\};$

5. 确定邻域函数值 N_c (协作过程)

6. 修正权系数

if $j \in N_c$ then

$$W_{ij}(t+1) = W_{ij}(t) + \eta(t) [X_{ik}(t) - W_{ij}(t)] \text{ else}$$

$$W_{ij}(t+1) = W_{ij}(t)$$

7. $n=n+1$ goto 2

4.2 SOM 图的标注方法

在构造好神经网络之后,即可进行预测。在进行若干次 SOM 计算后,生成一个二维映射图。其中每一点对应一个公司,但这只是一个数值的表示,因此,需要为每个点进行标识,标识方法为:

在输入向量 $X(x_1, x_2, \dots, x_n)$ 与每个输出层神经元节点的连接权值向量 W_j 的欧氏距离

$$d_{jk} = \|X_k - W_j\| = \sqrt{\sum_{i=1}^n [X_{ik}(t) - W_{ij}(t)]^2} \quad (2)$$

$j=1, 2, \dots, P$

之间进行比较,取与其最接近的节点,将此节点标注为该公司。

5 模型训练中迭代方式的改进

5.1 样本数据的随机选取

由于 SOM 分类与样本的输入顺序有关,在小样本训练的情况下,无疑样本数据要重复使用,这样有三种重复使用的方法可以选择:一是各样本数据按顺序循环输入;二是每次训练时均彻底改变样本数据次序,然后再输入;三是从样本集中随机选取样本数据。我们通过大量实验,表明第二种方式的训练效果较好,因此,在选择样本输入模型时,用程序对样本数据集进行了自动随机处理。

5.2 稀少样本的处理

在训练中,有些类别的样本很少,在这种情况下,代表这种类型的样本输入就不能得到充分训练,如果训练次数过少,会导致模型对这一类输入的分类结果产生偏差,模型预测精度下降;如果一味地加大整体训练次数,又会使得模型的训练时间过长。因此,本文利用以下方法对小概率样本进行选取,加强训练。

当模型经过一定的迭代之后,在初步形成的映射图上,有些输入模式在其迭代半径范围内所激活的神经元很少,如:图2所示为74个样本数据训练之后的映射图。其中的7号、8号样

