2018 年“花旗杯”金融创新应用大赛

机器学习报告

参赛题目：砺金--住房租赁资产证券化 REITs 平台

大赛队长：褚天硕

辅导老师：高 明、隋 聪

隶属学校：东北财经大学

目 录

1. 第一部分：背景概述

1.1 研究背景

砺金平台致力于向中小房地产企业提供信贷资金支持，将中小企业的担保资产打包后，成立 REITs 专项计划，并从中赚取佣金、服务费等。其中，对中小房地产企业是否具有还款能力而进行的信用等级评估，很大程度上决定了平台能否取得预期收益和能否稳健的运行。故根据机器学习中的有监督学习和无监督学习相结合的思想，本小组建立了一个针对分类结果的正确率而不断优化的复合神经网络模型，最后根据企业的信用等级而决定是否向企业发放贷款。

1.2 有监督学习

有监督学习是指输入的训练数据具有一定的指标，使得建立起来的机器学习模型在该指标的规范下对训练集实现高准确率的分类。砺金平台将中小企业的 13 项财务数据作为特征变量输入学习模型，是否为信用不良企业作为模型预期分类结果，将从新浪财经网、金融数据库等中提取出来的未知信用等级企业的相关财务信息作为样本，训练神经网络模型，得出分类结果，决定是否准许企业进入我们的平台。

2. 第二部分：模型选择

2.1 采用模型

通过对比各评级方法的优缺点，砺金平台最终选择采 SOM(Self-organizing feature Map)-LVQ(Learning Vector Quantization)复合神经网络模型对客户进行信用评级。首先，我们在参阅一些著名学者的文献的基础上，建立了评价指标体系，使用因子分析消除相关性，选取了具有代表性的因素进行模型的构建。其次，我们通过 MATLAB 进行模型构建，将经过筛选的数据输入到 SOM-LVQ 复合神经网络中进行训练和分类，最后对分类结果进行等级的划分，从而完成评级过程。

2.2 选择原因

（1）传统的要素分析法、比率分析法以及打分法都是对指标体系进行主观的分析，人为地对指标赋予权重；很明显，这些方法主观性太强，结果并不具备说服力。

（2）诸如 Logistic 回归模型、聚类分析 K 近邻判别等方法，其通常具有较为严格的前提使用条件，如样本数据需要服从正态分布等；而大量研究证实：

1）企业财务状况的研究可看作一系列独立变量的分类问题。

2）企业的财务状况的影响因素之间通常具有非线性的关系。

3）预测变量之间通常是高度相关的，并不能独立分析。以上种种因素进一步限制了模型的选择。

（3）我们所采集的企业样本所属类别预先未知，这也就意味着无法直接进行有指导的网络学习，如 BP 网络神经模型。而 SOM 网络神经是无指导学习模型，他能通过自动寻找样本的特点和本质规律，自适应地调整网络参数。同时，SOM 模型得出的权值，正好能满足 LVQ 模型有监督学习的前提要求。

（4）SOM 具有很强的自组织、自学习能力，能自动识别向量空间中的最明显特征，性能稳定，分类有效。LVQ 的优点在于算法简单、训练时间少，同时具有良好的识别效果。

（5）SOM 和 LVQ 神经网络模型都隶属于竞争网络的模式，在原理和算法上具有较为相似的特点，将二者有机结合，能更好的实现预测功能。

3. 第三部分：有监督学习神经网络和无监督学习神经网络

3.1 SOM 神经网络

SOM 神经网络全称是自组织映射神经网络(Self Organizing Maps)，可以对数据进行无监督学习聚类。SOM 神经网络只有两层网络结构，即输入层和竞争层/隐藏层：输入层对应输入的数据，竞争层/隐藏层通过对分类结果权值的自动化调整进一步优化竞争层的权重设置。在训练时采用“竞争学习”的方式，每个输入的样例在隐藏层中找到一个和它最匹配的节点，称为它的激活节点，又称为“winning neuron”。 紧接着用随机梯度下降法更新激活节点的参数。同时，

和激活节点临近的点也根据它们距离激活节点的远近而适当地更新参数。

3.1.1 操作步骤

（1）初始化：每个节点随机初始化自己的参数，每个节点的参数个数与 Input 的维度相同。

（2）对于每一个输入数据，找到与之最相配的节点。假设输入时为 D 维，即 X={x_i, i=1,...,D}，那么判别函数可以为欧几里得距离：

$$\mathrm{dj(x)} = \sum_{i=1}^{D} (x_i - w_{ji})^2$$

（3）找到激活节点 I(x)之后，我们也将更新和它临近的节点。令 Sij 表示节点 i 和 j 之间的距离，对 I(x)临近的节点，重新为其分配节点：

$$T_{j,I(x)} = \exp(-S_{j,I(x)}^2 / 2\sigma^2)$$

简单地说，临近的节点根据距离的远近，更新程度要逐步减少。

（4）下一步就是更新节点的参数。按照梯度下降法进行更新：

$$\Delta w_{ji} = \eta(t).T_{j,I(x)}(t).(x_i - w_{ji})$$

迭代，直到收敛。

（5）记录最后一次迭代的最大距离和最小距离。

3.2 LVQ 神经网络

LVQ（Learning Vector Quantization）神经网络是一种用于训练竞争的有监督学习方法的输入向前神经网络，其算法是从 Kohonen 竞争算法演化而来，能够实现有效的非线性分类。 LVQ 神经网络由三层神经元组成，即输入层、竞争层和线性输出层。如图 3-1 所示：

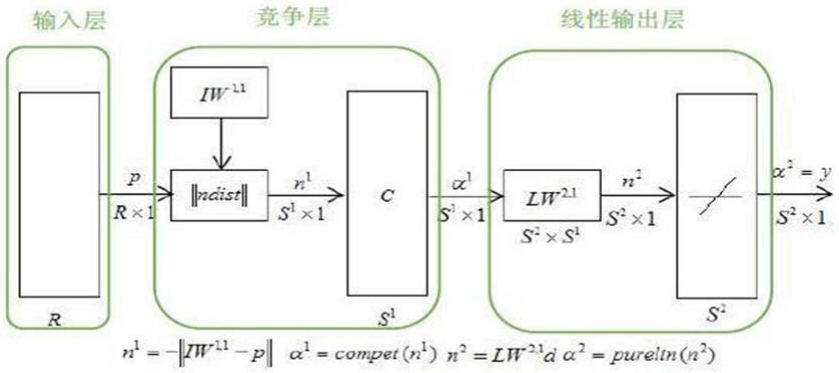

图 3-1

P 为 R 维的输入模式；

S^1为竞争层神经元个数；

$IW^{1,1}$为输入层与竞争层之间连接权系数矩阵；

n^1为竞争层神经元的输入；

α^1为竞争层神经元的输出；

$LM^{2,1}$为竞争层与线性输出层之间的连接权系数矩阵；

n^2为线性输出层神经元输入；

α^2为线性输出层神经元的输出。

4. 第四部分：具体实现过程

4.1 流程简介

首先将训练向量输入到 SOM 网络中进行迭代计算，通过设置步长，获得稳定的网络，之后将得到的训练样本分类结果作为 LVQ 网络的输入流，同时将 SOM 网络训练得到的网络参数作为 LVQ 网络的初始权值使用。整个过程通过对 SOM 与 LVQ 的有机结合，形成复合网络，在很大程度上避免了由于神经网络对初始权值赋值敏感而导致出现的训练结果的错误以及分类减少的误差，并减少了训练时间。同时，该复合神经网络比单一神经网络（如 BP 神经网络）使用起来更有效率，预测能力也得到很大的提高。

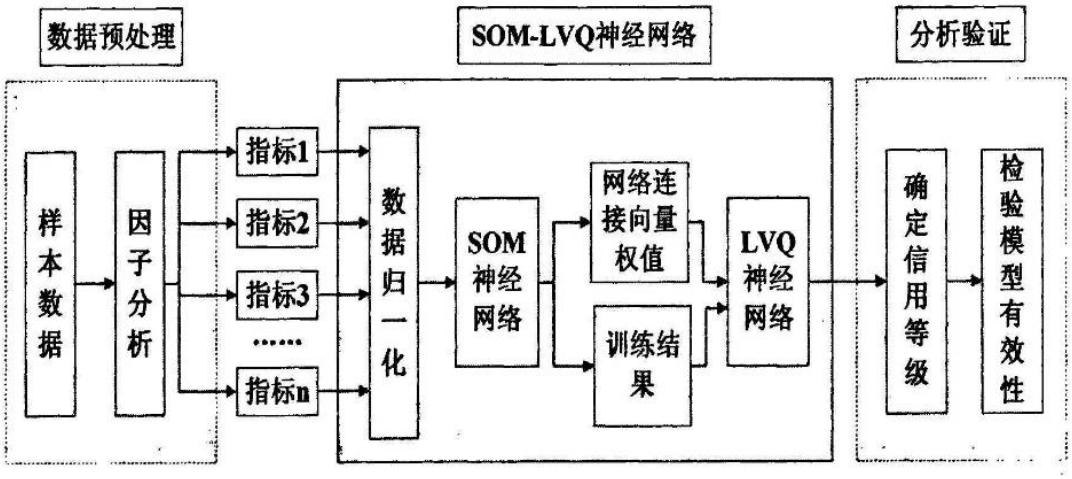

图 4-1

4.1.1 数据预处理

将采集来的样本企业数据输入到 SPSS 软件中，对偿债能力、营运能力、盈利能力、成长能力各自包括的指标分别进行因子分析，通过筛选，选取具有代表性的指标进入到 SOM-LVQ 神经网络模型中。通过筛选，我们选取净资产收益率、流动比率、速动比率、应收账款周转率、存货周转率、资产负债比率等 13 项指标作为最终的输入指标进入模型进行训练。

表 4-1

feature 1：基本每股收益	feature 2：每股净资产
feature 3：净资产收益率—加权平均	feature 4：扣除后每股收益
feature 5：流动比率	feature 6：速动比率
feature 7：应收账款周转率	feature 8：资产负债比率
feature 9：净利润率	feature 10：总资产报酬率
feature 11：存货周转率	feature 12：固定资产周转率
feature 13：总资产周转率	

4.1.2 SOM-LVQ 复合神经网络

由于财务指标变量为区间标度变量，企业规模指标为序数型变量；此外财务指标中的流动比率、存货周转率一般为正值，而资产利润率可正可负。因此，为

使不同指标之间具有可共度性，首先将数据进行归一化到[0，1]区间，然后将归一化的数据输入 SOM 模型中，进行训练，将经过训练后稳定有效的结果作为网络连接向量权值输入到 LVQ 模型中。LVQ 将输入的权值作为其初始网络连接向量权值，并通过 LVQ 神经网络的自我学习，以及 SOM 的训练结果作为监督，进行样本的优化训练，最终得到分类结果。

注：归一化公式：（采取最大最小线性转换法）

$$y = \frac{x - \text{MinValue}}{\text{MaxValue} - \text{MinValue}}$$

其中，MaxValue 和 MinValue 分别为样本中的最大值和最小值。

（1）我们使用 MATLAB 神经网络工具箱，利用 newsom 函数生成一个 SOM 神经网络，定义输入矩阵。其中，神经网络的拓扑函数，距离函数，分类阶段的学习速率，领域距离等参数均取默认值。

（2）将归一化后的样本输入 SOM 模型中，利用 train 函数和 sim 函数对神经网络进行训练。为保证实时观测训练结果，我们设置训练步数为 1000，2000，3000，4000，5000 和 6000，代码如下：

```
P = xlsread('C:\Users\dell\Desktop\huaqi\huizong');
n=13;
for i=1:n
     P(i,:)=(P(i,:)-min(P(i,:)))/(max(P(i,:))-min(P(i,:)));
end
P=P';    %%%取转置矩阵
PP=P(1:1:57,:)';
net=newsom(minmax(PP),[1 4]);
a=[1000 2000 3000 4000 5000 6000];
for i=1:6
    net.trainParam.epochs=a(i);
    net=train(net,PP);
    y=sim(net,PP);    %进行仿真
    yc=vec2ind(y);    %建立索引
end
```

```
pt=AA;
C=yc;
T=ind2vec(C);
n1=1;
n2=1;n3=1;n4=1;
n5=1;n6=1;n7=1;
n8=1;n9=1;n10=1;
n11=1;n12=1;n13=1;
a1=[];a2=[];a3=[];
a4=[];a5=[];a6=[];
a7=[];a8=[];a9=[];
a10=[];a11=[];
a12=[];a13=[];
for i=1:100
    if C(i)==1
        a1(n1)=i;
        n1=n1+1;
    end
    if C(i)==2
        a2(n2)=i;
        n2=n2+1;
    end
    if C(i)==3
        a3(n3)=i;
        n3=n3+1;
```

（3）经过实验发现，当神经网络的训练步数为 6000 时，分类结果达到稳定，权值不再发生变化。

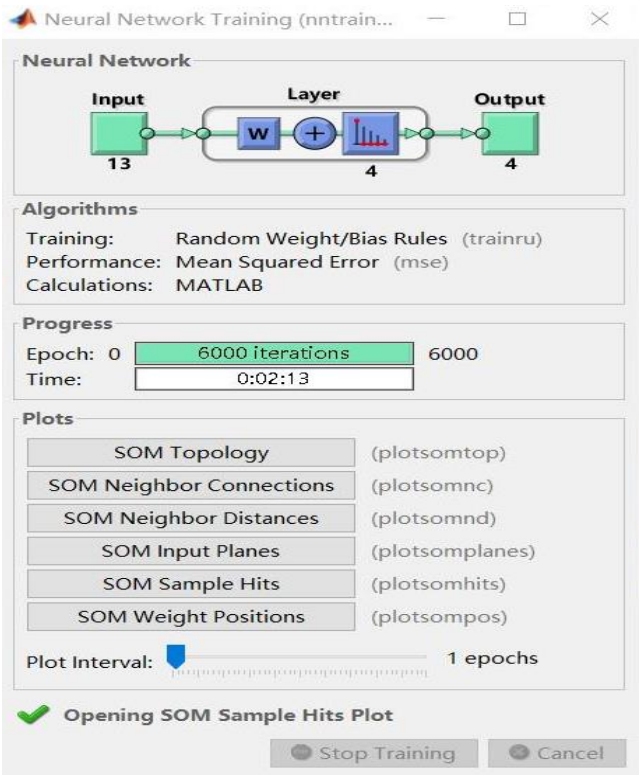

图 4-2

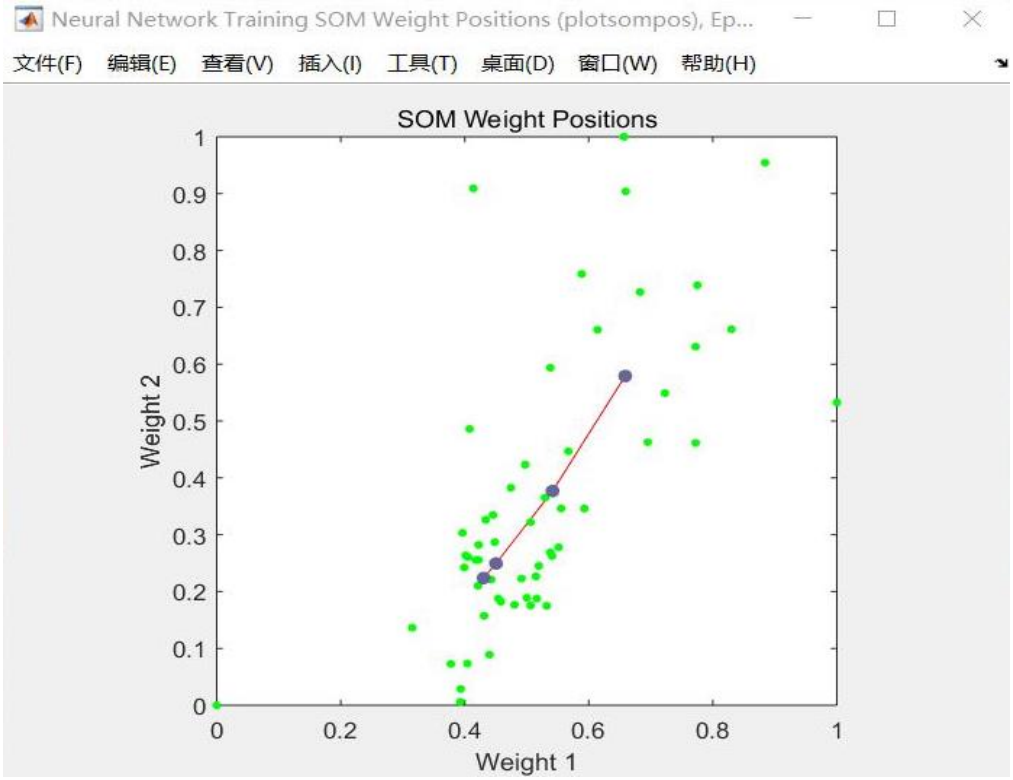

图 4-3

（4）接着，使用 newlvq 函数生成一个 LVQ 神经网络，将 SOM 的训练结果作为LVQ网络的输入向量，同样，对LVQ神经网络的拓扑函数等参数也均取默认值。

（5）使用函数 train 和 sim 对神经网络进行仿真训练，设置训练步数为 20，40，60，80，100，代码如下：

```
Pt=PP;
C=yc;
T=ind2vec(C);
n1=1;
n2=1;n3=1;n4=1;

a1=[];a2=[];a3=[];
a4=[];
for i=1:57
    if C(i)==1
        a1(n1)=i;
        n1=n1+1;
    end
    if C(i)==2
        a2(n2)=i;
        n2=n2+1;
    end
    if C(i)==3
        a3(n3)=i;
        n3=n3+1;
    end
    if C(i)==4
        a4(n4)=i;
        n4=n4+1;
    end
end
```

```
B=[(n1-1);(n2-1);(n3-1);(n4-1)]'/57;
lvqnet=newlvq(minmax(Pt),4,B);
lvqnet.IW{1,1}=net.IW{1,1};
a=[20 40 60 80 100];
for i=1:5
    lvqnet.trainParam.epochs=a(i);
    lvqnet=train(lvqnet,Pt,T);
    yy=sim(lvqnet,Pt);
    C=vec2ind(yy);
end

```

（6）分类结果显示：

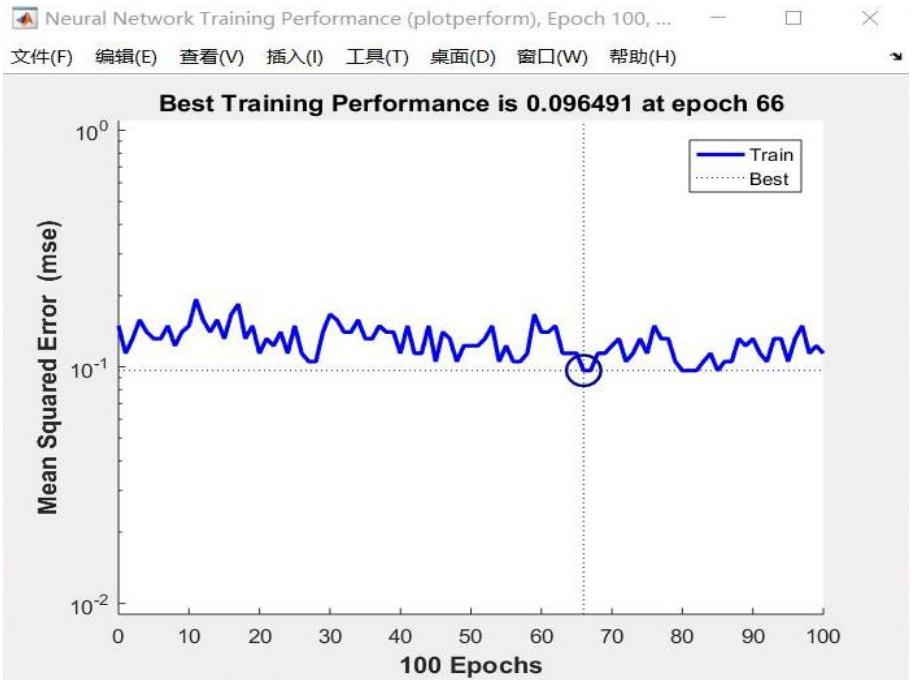

图 4-4

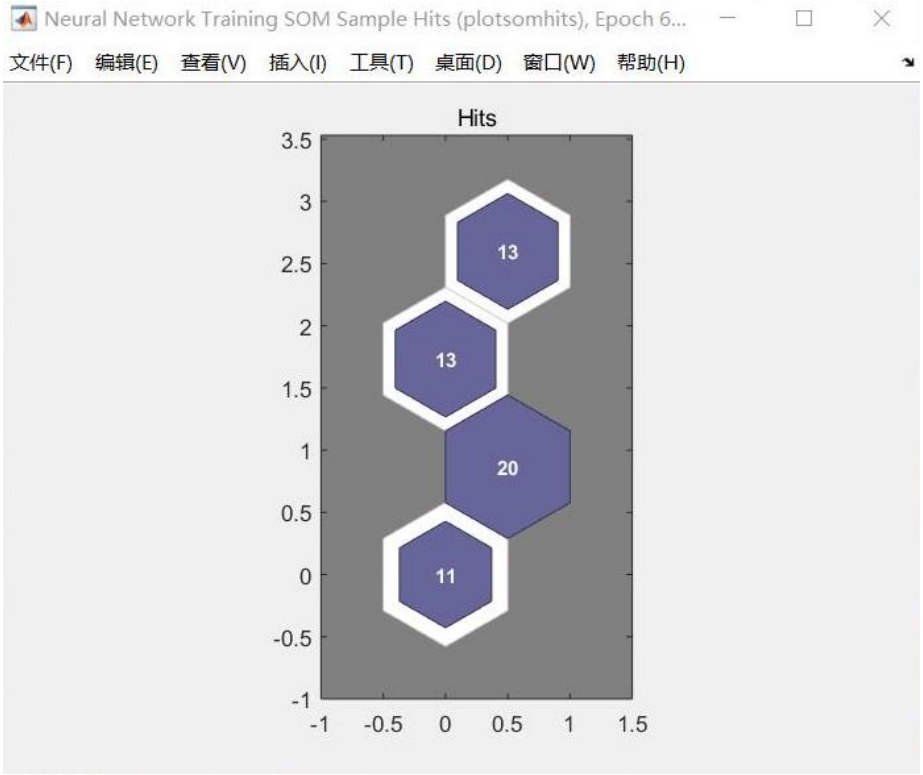

图 4-5

结果如上所示，共输入了 57 家样本企业数据进行训练，最终得到 4 类结果。

注：57 家训练样本数据如下表 4-2：

表 4-2

训练企业序号	企业名称	训练企业序号	企业名称
1	万科	30	绿地
2	金地	31	保利
3	荣丰控股	32	新华联
4	荣安地产	33	深大通
5	华远地产	34	金融街
6	泛海建设	35	沙河实业
7	渝开发	36	万泽股份
8	北京城建	37	苏州高新
9	阳光城	38	华侨城
10	中粮	39	顺发恒业
11	银亿股份	40	海航基础
12	泰禾集团	41	卧龙地产
13	华业资本	42	中天金融
14	大名城	43	上海万业
15	栖霞建设	44	三湘印象
16	华丽家族	45	大港股份
17	上海临港	46	新黄浦
18	浙江广厦	47	卧龙地产
19	大东海	48	华天酒店
20	锦江股份	49	荣盛股份
21	联络互动	50	全新好
22	京投发展	51	阳光股份
23	*ST 天业	52	ST 岩石
24	东洋	53	京汉股份
25	苏宁环球	54	雷伊
26	京能置业	55	首旅酒店
27	光大嘉宝	56	皇庭国际

28	世联行	57	西藏城投
29	金灵通		

4.2 分析验证

在第二步分类结果的基础上，对结果划分信用等级，同时收集已知等级的企业样本数据作为验证集进行验证，确保分类结果的有效性。验证结果显示如下：

表 4-3

分类	所属公司
1	16，27，64，73，85，90，93，95，99
2	29，35，26，66，91
3	11，41，37，46，58，70
4	2，4，10，15，40，38

4.3 模型评价

通过验证集的验证，我们发现分类结果的正确率达到 80%，特别是对信用等级很高的企业和信用等级很低的企业能够进行准确的判断，对于中间等级的企业评级有较小的偏差，对信用评估影响微弱。由此可见，我们的 SOM-LVQ 复合神经网络是有效的。