自组织映射聚类算法在电信客户细分中的应用①

吴春旭　鲍满园　苟清龙 (中国科学技术大学 管理学院 安徽 合肥 230026)

摘　要： 将自组织映射SOM（Self Organization Map）聚类算法应用于电信客户细分，并与采用K-means聚类算法得到的结果进行比较。实验表明，SOM可以有效的进行电信客户细分且聚类效果较优，但需付出训练时间的代价。同时对两种算法的复杂度、误差等进行了分析。

关键词： 自组织映射；神经网络；电信；聚类；客户细分

Application of Self Organization Map to Classification of the Telecommunication Company

WU Chun-Xu，BAO Man-Yuan，GOU Qing-Long

(Department of Management Science, University of Science and Technology of China, Hefei 230026, China)

Abstract: This paper applies the SOM neural network to the customer segmentation of the telecommunication company, and compares the results of the K-means clustering algorithm and the SOM. The experiment indicates that the SOM is effective to classificate and the cluster effect is better when the data assemble is large, but it takes the training time. At the same time, it analyses the algorithm complexity and the errors of the two algorithms.

Keywords: SOM; neural network; telecommunication; clustering; customer classification

1　引言

随着市场经济的发展，企业为了占有更多的市场份额，越来越重视客户的细分。客户细分是根据消费者购买行为的差异性，把消费群体划分为相似性购买群体的过程。客户细分可以帮助企业找到一些高价值模式，没有客户细分，则企业将被很多的低价值模式所困扰[1]。

对于电信企业来说，不同的客户群体具有不同的内在价值，根据客户的消费数据，把客户分为不同的类或簇，从而发现同一类或同一簇的客户群的消费特点，并据此制定差别化服务政策。客户细分的实现方法也有多种，常见的有蚁群算法，K 均值算法等。分类方法通常有基于分区和基于模型的分类方法。K 均值算法就是一种典型的基于分区的聚类算法，而 SOM(自组织映射神经网络)是一种基于模型的分类算法。

RFM 模型[1]是一种有效的客户细分模型。林盛等学者[2]将 RFM 模型应用电信客户的细分，取得了一定的成果，本文利用这一模型，采用自组织映射(SOM: Self Organization Map)聚类算法，对电信客户进行细分，得出具有不同客户价值[2]的客户群，并将结果与采用 K 均值算法得到的结果进行比较，为了直观反映聚类的效果，采用了三维原数据的聚类。

2　电信业RFM模型和SOM 算法

2.1 用于电信客户细分的 RFM 模型

RFM 模型是企业客户分类的主要方法之一，它使用的三个指标是近度 R(Recency)、频度 F(Frequency)、值度 M(Monentary)。通常不能直接将 RFM 模型运用于电信行业的客户细分，而是从客户交费角度来建立电信业客户细分的 RFM 模型(其模型指标与

① 基金项目:安徽省自然科学基金(090416240);高等学校优秀人才基金(2009SQRS001ZD)
收稿时间:2009-11-25;收到修改稿时间:2009-12-30

传统的 RFM 指标含义比较如表 1 所示。)[2]

表 1 一般的 RFM 模型与电信业 RFM 模型的含义比较

模型	R	F	M
一般的 RFM 模型	客户最近一次购买距离分析点的时间	客户一定时期内购买该企业产品的次数	客户一定时期内购买该企业产品的总金额
电信业 RFM 模型	客户最后一次交费距离分析点的时间	客户一定时期内交费的次数	客户一定时期话费的总额

2.2 自组织映射(SOM)神经网络算法

自组织特征映射神经网络。(Self-organizing Feature Maps)简称 SOFM 或者 SOM，是由芬兰赫尔辛基大学神经网络专家 Kohonen 教授在 1981 年提出的[3]。自组织映射网络是一种无导师聚类算法，即可以对未知聚类结果的数据集进行聚类。SOM 网络是由输入层和输出层(也叫竞争层)组成的单层神经网络。输入层有 n 个神经元，输入层神经元的个数由输入向量的分量的个数决定。输出层有 m 个神经元，一般按二维的形式排成节点矩阵。输入层的神经元和输出层的神经元都有权值连接，输出层节点相互之间也可能有局部权值连接。其网络模型如图 1 所示。

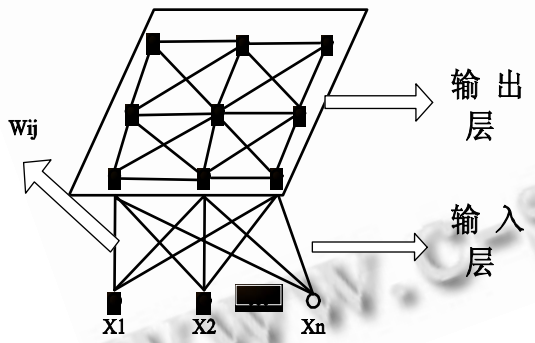

图 1　SOM 网络结构

输出层的第 i 个神经元与输入层的神经元连接权值用 $W_{ji}=\{W_{1i},W_{2i},\cdots,W_{ni}\}$表示， SOM 网络通过对输入模式的反复学习，在输出层将获胜的神经元表示出来，获胜的神经元与输入模式有最小的欧式距离，即获胜神经元 i*的获胜条件为：

$$\| X_S - W_{ji^*} \| = \min \| W_S - W_{ji} \| \qquad i \in m \ , \ j \in n$$

式中 $X_s=\{X_{s1},X_{s2},\cdots,X_{sn}\}$表示一个输入模式，而获胜的神经元 i*依据 Kohonen 学习规则只对其周围一个小邻域 $N_i^*(d)$进行权值修正。$N_i^*(d)$表示获胜神经元 i*周围以 d 为半径的区域。

$$N_{i^*}(d) = \exp(-\frac{|q-i^*|^2}{2\sigma^2}) \tag{1}$$

或

$$N_{i^*}(d) = \exp(-\frac{|q-i^*|^2}{\sigma^2}) \tag{2}$$

$$\sigma = \sigma(n) = \sigma_0 \exp(-\frac{n}{\tau}) \ , \ n=0,1,2,\cdots$$

其中 σ_0 是 σ 的初值，是时间常数，n 是迭代次数，q 是 $N_i^*(d)$内的其他神经元，$|q-i^*|$表示邻域内神经元与获胜神经元的距离，随时间的增加而变小 $N_i^*(d)$的大小随时间而收缩。获胜的神经元及 $N_i^*(d)$范围内的神经元的权重的调整规则为：

$$W_{ji}(t+1) = W_{ji}(t) + \eta(t)(X_s - W_{ji}(t)) \tag{3}$$

其中 $\eta(t) = \eta_0 \exp(-\frac{n}{\tau_2})$ n=0,1,2,…，是学习速率(增益函数)，η_0 是初始学习率，τ_2 是另外的一个时间常数，n 是迭代次数，$\eta(t)$ 随着时间的增加而减小。

通过以上的方式不断的调整各神经元的权值，就会形成输入空间到神经元节点集的一个映射，而且节点之间形成一种特定的位置关系，使输入空间中相近的样本映射到节点平面上相邻的节点上。SOM 网络通过自学习形成了一个对输入空间的内部表示，这种表示一方面反映了原空间样本的密度分布，另一方面保持了原空间样本之间的拓扑关系，这些信息被高度压缩到一个简单的有限个节点的平面上，有某些样本在空间中相对集中地聚在一起，形成了聚类。

3　运用在电信客户细分中的SOM算法

电信客户 RFM 模型的各维度的量纲并不一样，所以为了克服量纲不同的影响，可以采用归一化来解决。对于和客户价值正相关的数据式采用：$x'=(x-x^s)/(x^l-x^s)$，对于和客户价值反相关的数据式采用：$x'=(x^l-x)/(x^l-x^s)$。其中 x’为标准化过后的值，x 为原值，x^s 为相应指标最小的值，x^l 为相应指标最大的值，标准化过后的数据其值在[0,1]之间。SOM 在电信客户细分中运用的思路如下。

①以电信客户的 R、F、M 作为划分客户的指标，3 个属性值的最大的类别为 2*2*2=8 类，为了包含所有的类别模式，可以假设神经网络节点的个数为 9 个(3*3)[4]，表示最多可以分为 9 类。

②为了得到较好的聚类效果，我们采用如图 2 所示的六角形结构的 SOM 网络拓扑结构[4]。将初始权值 W_{ji} 赋予较小的随机初始值。设置一个相对较大的初始邻域 N_d，设置网络循环次数[5]T=500。

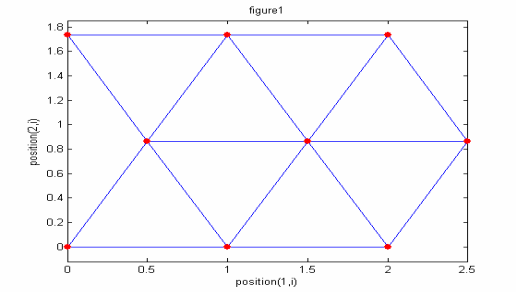

图 2　网络拓扑结构

③给出一个输入模式 $X^s=\{X^s_1,X^s_2,\cdots,X^s_n\}$(n 表示输入数据的维，s=1，2，…)输入网络，让网络学习输入模式。

④计算输入的模式和所有的神经元的欧式距离 $d_i=\sqrt{\sum_{i=1}^{n}(x_s-w_{ji})^2}$，并选取和 X_s 距离最小的神经元 i* 为获胜神经元。

⑤改正神经元 i*及其邻域节点的连接权值用(3)计算。邻域函数采用(2)式表示。

⑥继续输入新的模式，返回到④，直到全部模式输入到网络。

⑦等到网络训练步数达到最大训练步数，向训练好的网络输入一组测试数据集 $X_c=\{X_{c1}, X_{c2},\cdots,X_{ck}\}$，看看聚类的效果，以此来确定网络是否过度的适应了训练数据集[6]。如果发现网络的聚类效果很差，则返回②，重新设定训练的参数。否则网络已经完成训练。

⑧计算整个客户数据集的 R、F、M 的平均值和每个簇的 R、F、M 的平均值，并进行比较。

⑨根据 R、F、M 各自的权值[2]：$[w_R, w_F, w_M]$ = [0.221，0.341，0.439]，计算每一簇的客户的 RFM 的加权值，再根据加权结果重新进行分类。

通过以上的思路就可以得到按照客户价值的大小而聚类的客户群，针对不同的客户价值采取不同的服务策略。

4　仿真结果与分析

本文利用 MATLAB 工具，通过随机产生 1000 组[0，1]均匀分布的数据，设此数据集为经过标准化后的的数据。应用训练好的 SOM 网络，对数据进行聚类，得到 SOM 训练后的网络权值(T=500,初始的学习率为 0.1)。可以看出 SOM 网络把数据初步归为 9 类。用计算机计算出 1000 组数据的 R、F、M 的各自的平均数，并且将每一簇的中心和总体平均数比较，较之总均值大的用“↑”标记，反之则用“↓”。得到表 2。

表 2　SOM 的聚类结果

簇编号	近度	频度	值域	比较结果
1	0.20466	0.66902	0.54372	↓↑↑
2	0.37374	0.68441	0.38423	↓↑↓
3	0.60876	0.54533	0.2256	↑↑↓
4	0.44977	0.64294	0.63011	↓↑↑
5	0.53423	0.42397	0.44779	↑↓↓
6	0.60704	0.28425	0.33983	↑↓↓
7	0.70207	0.65436	0.76391	↑↑↑
8	0.50883	0.52328	0.73763	↓↑↑
9	0.52351	0.17943	0.56742	↑↓↑
总均值	0.5138	0.4948	0.5128	

表 2 显示：第一簇、第四簇和第八簇可以归为同一类，第五簇和第六簇也可以归为同一类。其各类的数目如图 3。

由于 R、F、M 各自的权重不一样，所以以上的聚类结果并不是最后的结果。根据 R、F、M 各自的权重，计算各簇的权重加权值，并将此值视为客户价值，其大小代表着客户价值的大小。根据客户价值的值，排列出各客户群价值的大小。

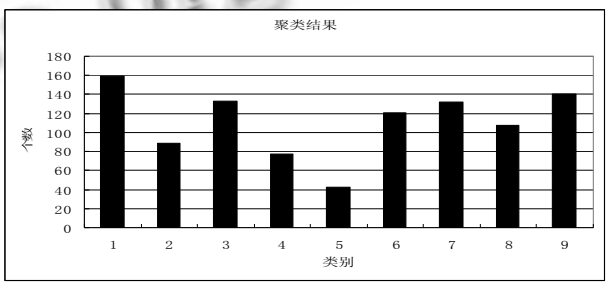

图 3　SOM 聚类结果直方图

第一簇到第九簇的权重加权值为：0.5121，0.4847，0.4195，0.5953，0.4592，0.3803，0.7137，0.6147，0.4260。得到的各类对应的权值按照大小顺序为：簇 7>簇 8>簇 4>簇 1>簇 2>簇 9>簇 5>簇 3>簇 6，综合权重和表 2 得到图 4。(图中实线的折线所示。)

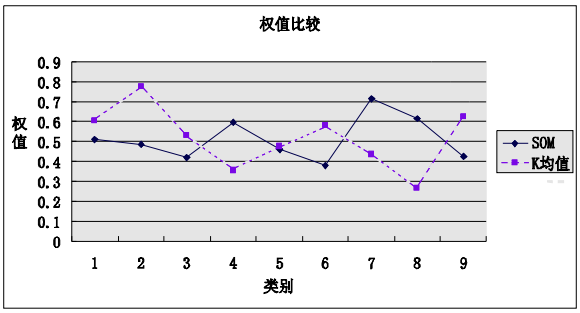

图 4　聚类的权值比较

从图 4 表示按照客户价值可以分为五类,具体如下。

(1)第一类是簇 7 即权重是 0.7137，这类客户群可以被认为不仅频繁的缴费而且花费较多，是电信企业应该保持的客户群。

(2)第二类是簇 4 和簇 8，其权重很相近。此类客户群与企业接触较为频繁，交易额较大。是电信企业重要的发展客户群。

(3)第三类是簇 1 和第 2，此为企业的一般重要的客户。

(4)第四类是簇 3、簇 5 和簇 9，此类客户群可以看做是一个新的客户群。可以培养这类客户群成为重要的客户群。

(5)第五类是簇 6，这类客户的价值较小。

电信企业应该优先针对用户价值大的类别的特点推出适合该类的服务策略，对于某些价值处于中间的类可以培养其成为价值更大的类。

本文还用 K 均值算法进行了一次实验，结果如表 3 所示。

表 3　K 均值的聚类结果

簇编号	近度	频度	值域	比较结果
1	0.77202	0.56214	0.55219	↑↑↑
2	0.7081	0.79863	0.79282	↑↑↑
3	0.75017	0.78126	0.21847	↑↑↓
4	0.77464	0.21405	0.25214	↑↓↓
5	0.26546	0.21444	0.78546	↓↓↑
6	0.76639	0.18395	0.78927	↑↓↑
7	0.24874	0.76272	0.27852	↓↑↓
8	0.26452	0.26119	0.26753	↓↓↓
9	0.21191	0.71162	0.76914	↓↑↑
总均值	0.5138	0.4948	0.5128	

从表 3 可以看出簇 1 和簇 2 是同一类，结合给出的权重,得到第一簇到第九簇的权重加权分别为：0.6047，0.7769,0.5281,0.3549,0.4766,0.5786,0.4373,0.2650，0.6271。

SOM 和 K 均值算法聚类结果有着明显的不同，本文从以下的几个方面进行了一些比较。

①时间

K-means 方法实验的时间远远小于 SOM 的时间，但是 SOM 的时间长是因为其需要训练网络，一旦网络训练结束后，其聚类的时间也较短。

②权重的比较

从图 5 可以很明确的看出，根据客户价值的高低基本可以划分为五类客户群，两种算法都可以大概分出五类。

③聚类结果的分析

我们用 K 均值聚类得到各类的数目如图 5 所示。

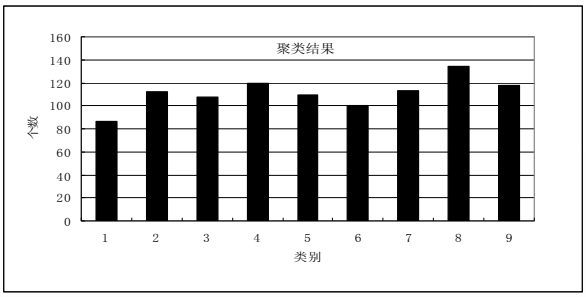

图 5　K 均值聚类结果直方图

K 均值得到的各类对应的客户价值的排列顺序为：簇 2>簇 9>簇 1>簇 6>簇 3>簇 5>簇 7>簇 4>簇 8。实验得到：在小数据集的情况下，K 均值的误差要明显小于 SOM 的误差，但是当面对大数据集的时候，SOM 的表现要好于 K 均值的表现。此外当电信企业需要进行预测的时候，K 均值显得无能为力，因为 K 均值只要加入一条新的记录，其整个的分类结果就有可能产生变化，但是 SOM 则可以利用训练后的网络很轻易的预测。

对于电信客户的 RFM 模型，SOM 算法较适合客户细分。尽管在小数据集的时候 K 均值表现的要比 SOM 好，但是电信客户的数据量是巨大的，且 SOM 还可以预测客户的消费行为，但是 SOM 算法训练网络花费的时间也是一个要考虑的因素。

5　结语

本文把 SOM 算法运用在电信客户细分的 RFM 模型中，根据客户价值的大小将电信客户划分为若

干不同的类别，实验结果表明 SOM 方法可以有效的进行电信客户细分。同时本文还将 SOM 和 K 均值算法进行比较，得到两种算法的不同的聚类效果，并对这两种方法的聚类效果进行分析，发现这两种算法各有优劣。SOM 算法是一种无导师学习的算法，不仅可以用来聚类还可以用来预测。但是其还不成熟，在运用的时候有许多的问题要解决。在今后的客户细分运用中仍然有很多工作去做。如：①怎样避免网络训练时出现“死神经元”；②如何确定网络训练的相对较合适的训练步数；③可否设置出一个客户价值的合理的参考阈值。

参考文献

1 Chen YL. Kuo MH, Wu SY, Tang K. Discovering recency, frequency, and monetary(RFM)sequential patterns from customers' purchasing data. Electronic Commerce Research and Applications, 2009(8):241－251.

2 林盛,肖旭.基于 RFM 的电信客户市场细分方法.哈尔滨工业大学学报, 2006(5):758－760.

3 Kohonen T. Self-organized formation of topologically correct feature maps, Biological Cybernetics, 1982, 43(1):59－69.

4 Budayan C, Dikmen I, Birgonul MT. Comparing the performance of traditional cluster analysis, self-or ganizing maps and fuzzy C-means method for strategic grouping. Expert Systerms With Applications, 2009,36: 11772－11781.

5 杨占华,杨燕.SOM 神经网络算法的研究与进展.计算机工程, 2006,(8):201－228.

6 董长虹.MATLAB 神经网络与应用.第 2 版,北京:国防工业出版社, 2007.147－170.