基于主成分—聚类分析方法的客户分类研究

王建民[1], 王传旭[2]

(1.安徽理工大学 经济管理与社会科学系, 安徽 淮南 232001;
2. 淮南师范学院, 安徽 淮南 232001)

[摘 要] 客户分类比较理想的依据是客户对企业的利润贡献度。文章通过构建企业与客户交易记录的原始数据矩阵,应用主成分—聚类分析的方法来定量地研究客户对企业的利润贡献度,进行客户分类工作。

[关键词] 客户分类;利润贡献;主成分分析;聚类分析
[中图分类号] F274 [文献标识码]A [文章编号] 1009-9530(2006)03-0077-03

1 引言

客户关系管理(CRM)的效益在一定程度上建立在对客户分类的科学性和艺术性上。客户分类是根据企业发展的需要,依据某些指标体系和方法对企业的客户进行的一系列系统的分析、归类和总结,从而服务于客户关系管理。

目前,国内、外众多的学者对客户分类问题进行了研究,取得了很多成果。本人提出一种基于主成分—聚类分析的方法,计算客户对企业的利润贡献度,来研究客户分类问题。

本文拟解决的客户分类问题描述如下:在客户关系管理(CRM)中,某企业拥有若干个客户,每个客户在与企业的交易中有交易总额、交易数量、购买价格、交易频率、衍生交易数量、客户接触成本等个指标,通过对客户进行主成分——聚类分析,根据客户对企业利润贡献度,进行客户分类。

2 主成分—聚类分析

2.1 主成分分析法

主成分分析(principal component analysis),是研究如何把多个相关变量综合成一个或少数几个综合指标,而这一个或少数几个综合指标又能最大程度地反映原来变量信息的一种多元统计分析方法。

主成分分析法的核心思想是降维。通过主成分分析可以把具有相关关系的多个因子转化为一组相互独立的少数几个综合因子。这些综合因子将各原始指标因子中重叠的信息去掉,达到仅包含各原始指标间明显的差异、并且反映原始指标因子的主要信息之目的。即在不改变原始数据所提供的信息的基础上更集中、更典型地显示出研究对象的本质特征。主成分分析法在社会和经济统计研究中的应用十分广泛。

2.2 聚类分析法

聚类分析(cluster analysis),是一种数据简化技术,它把基于相似数据特征的变量组合在一起,成为一个类别。通过聚类分析,可以把分类对象按一定规则分成若干类,这些类不是事先给定的,而是根据数据的特征确定的。同一类对象之间在某种意义上差别较小,不同类对象之间的差别较大。也就是把一个集合分成若干个子集。

聚类分析是研究'物以类聚'的一种多元统计方法。在实际应用中,通过对相对比较完整的原始数据记录进行统计和分析,挖掘出有用的信息,合理有效地提高信息的利用率。这种技术对发现基于相似特征(如利润贡献水平、购买行为或购买习惯)的客户分类非常有用。

2.3 应用分析

在客户分类中,由于涉及盈利情况、交易总额、交易价格、交易频率、衍生交易数量、客户接触成本、客户诚信度、客户所处生命周期等诸多指标,各项指标之间往往具有一定的相关性,对运算和分析问题带来了不必要的麻烦,使企业在客户分类中,不能很好的抓住主要矛盾,看清各客户对企业的利润贡献情况,分出主要客户。而人为地选择分类的

[收稿日期] 2006-04-15
[作者简介] 王建民(1978—),男,河南泌阳人,安徽理工大学管理工程与科学硕士研究生,研究方向:管理工程与科学。王传旭(1955—),男,安徽淮南人,淮南师范学院教授,硕士生导师。

指标变量会带有主观意识，影响到分类的科学性。所以必须对所考虑的众多指标变量，用数理统计的方法，经过正交化处理，变成一些相互独立、为数较少的综合变量，再以这些综合变量作为聚类分区的新的数值数据。

主成分分析法(PCA)、聚类分析法(CA)为实现这一思想提供了十分有效的数学方法。应用过程是，首先确立指标体系，建立原始数据矩阵，然后用主成分分析法对原始数据进行筛选，根据筛选出的主成分，用聚类分析法得出客户分类方案。

3 计算过程

3.1 构建原始数据矩阵

(1)原始数据矩阵的构建

设某企业根据交易记录，有 n 个客户的样本 x_i $(i=1,2,\cdots,n)$，每一个客户样本有盈利情况，交易总额、交易价格、交易频率、衍生交易数量、客户接触成本、客户诚信度、客户所处生命周期等 m 个指标因子 $x_j(j=1,2,\cdots,m)$，所得观测值为 $x_{ij}(i=1,2,\cdots,n;\ j=1,2,\cdots,m)$，构成原始数据矩阵 $X=(x_{ij})_{m\times n}$

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{11} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

接下来，应用数理统计的方法，对众多具有一定相关性的原始指标数据，进行线性组合变换，重新组合为新的综合指标，并在所有的组合中选出包括信息量较多的前 r 个综合指标，这 r 个指标的累计方差贡献率一般以达到 85%或以上为原则。这 r 个主成分不仅保留了原始变量的主要信息，而且彼此无关，同时又比原始变量具有某些更优越的性质，这使得我们在研究复杂问题时，容易抓住主要矛盾。

3.2 计算主成分

(2) 将原始数据标准化

由于原始记录的各指标数量级和量纲不同，造成数值差别悬殊，为排除这些影响，使各种评价指标具有可比性，采用如下公式（标准差标准化方法），对原始数据进行标准化处理：

$$x_{ij}^{*}=\frac{x_{ij}-\overline{x_j}}{s_j} \qquad i=1,2,\cdots,n;\quad j=1,2,\cdots,m$$

其中，$\overline{x_j}$和 s_j^2分别是第 j 个指标的样本均值和样本方差，且

$$\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$$

$$s_j^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2$$

(3)计算相关关系矩阵

在标准化数据矩阵 $X^{*}=(x_{ij}^{*})$ 的基础上，计算原始指标的相关系数矩阵 $R=(r_{ij})_{m\times m}$，其中，r_{ij} 是 x_i 指标因子与 x_j 指标因子的相关系数，且

$$r_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}x_{ki}^{*}x_{kj}^{*}=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2(x_{kj}-\overline{x_j})^2}}$$

$i,j=1,2,\cdots,m$

(4)求解相关矩阵的特征根和特征向量

用 Jacobi 法对相关矩阵 R 作正交变换，化为对角型矩阵。即存在正交阵 Q 使

$$Q^{T}RQ=\lambda_1\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{bmatrix}$$

其中，$\lambda_1,\lambda_2,\cdots,\lambda_m$为 R 的 m个特征根。不妨设 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m$，对应的标准正交特征向量，为

$$a_j=(a_{1j},a_{2j},\cdots,a_{nj}) \qquad j=1,2,\cdots,m$$

则

$$a_i a_j^{'}=\begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$$

而特征根 λ_j 为主成分 Y_j 的方差，方差越大对总变异的贡献越大；特征向量 a_j 则是主成分 Y_j 的线性表达式中原始指标(已标准化)的组合系数。即

$$Y_j=\sum_{k=1}^{m}a_{kj}^{*}x_k^{*} \qquad j=1,2,\cdots,m$$

(5)计算方差贡献率和累计方差贡献率

方差贡献率是用来表明主成分综合原始数据能力强弱的指标，第 j 个主成分 Y_j的贡献率为：$e_j=\frac{\lambda_j}{\sum_{k=1}^{n}\lambda_k}$.它表示主成分 Y_j所反映的信息量的大小。

累计方差贡献率，是表示前个主成分所提取的原始数据信息量的比重，$G(r)=\sum_{k=1}^{r}e_k$

(6)提取主成分

在已确定的全部 m个主成分中合理选择前 r 个来实现最终的评价分析。实际应用中，确定 r 的值通常要使 G(r)达到 70%-85%或以上为原则。根据经验，r 的值往往不超过 3。

(7)计算各指标因子负荷量

因子负荷量是刻画主成分经济意义的重要指标，反映所取主成分与各原始指标之间的相关关系，其绝对值的大小是对主成分进行经济解释的重要依据。它是各特征值的方根与其对应的特征向量的乘积($\sqrt{\lambda_j}\,a_j$)。再算出每个指标在各主成分上负荷量的平方和($\sum_{j=1}^{r}\lambda_j a_{ij}^2$ $i=1,2,\cdots,m$)，即公因子方差。它反映各指标对选出的主成分所起的作用，即原始数据的重要程度。对主成分进行解释时结合定性分析进行，将所选主成分与工作和定性分析中的直觉加以对照比较还可以检验计算结果的

合理性及恰当程度。

(8) 计算 n 个调查客户的样本在前 r 个主成分上的得分：

主成分得分是原始数据(已标准化)在主成分所定义的新坐标系中的新数据。即

$$Y_j = \sum_{k=1}^{m} a_{kj}^{*} x_k^{*} \qquad j=1, 2, \cdots, r$$

3.3 系统聚类

(9) 对经过主成分分析得到的新数据($Y_1, Y_2, \cdots, Y_r$)进行系统聚类分析：

1) 规定样本之间的距离 d_{ij} 和类与类之间的距离 D_{KL}

本文分别采用欧氏距离公式计算 d_{ij} 和类平均法计算 D_{KL}. 即

$$d_{ij} = \sqrt{\sum_{k=1}^{r} |x_{ik} - x_{jk}|^2} \qquad (1)$$

$$D_{KL}^2 = \frac{1}{n_k \cdot n_l} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} d_{ij}^2 \qquad (2)$$

式(1) 中，i 和 j 是样本序号，k 是主成分序号。式(2) 中，n_k 和 n_l 分别是类 G_K 和 G_L 的样本个数。类平均法较好地利用了所有样本之间的信息，在很多情况下它被认为是一种比较好的系统聚类方法。

2) 计算类与类之间的距离矩阵 D

开始时，n 个调查客户的样本各成一类，组成 n 类：$G_1, G_2, \cdots, G_n$。故开始时，类与类之间的距离与样本之间的距离相同，即

$$D=(d_{ij})_{n\times n}$$

3) 进行系统聚类

将 D 中最小元素对应的类合并，然后按照公式(2) 重新计算新类与相邻类的距离。如果全部的类都已成一类，则过程终止，否则回到 3).

4) 根据需要确定阈值，将各调查客户(即各样本)划分为 s 个类。

(10) 计算各个类别的综合得分 Z:

1) 计算每个客户的综合得分 $Z_{1-r} = \sum_{k=1}^{r} e_k \times Y_K$. 其中 e_k 是方差贡献率。

2) 计算各个客户类别的综合得分 $Z=(Z_{(1)}, Z_{(2)}, \cdots, Z_{(s)})$. 其中，$Z_{(i)}$ ($i=1, 2, \cdots, s$) 是第 i 类中所有客户综合得分的平均值。

据此，我们可以根据每个客户类别的综合得分情况，将企业的客户划分为黄金客户、白银客户、普通客户、危险客户、淘汰客户等类别，并根据每个客户的综合得分情况，将该客户归入其所应在的客户类别。从而实现对客户的科学分类。

计算过程进行完毕。

4　结束语

本文基于主成分—聚类分析的方法来研究客户分类问题。具体的计算过程我们可以借助软件 SPSS11.0 来实现。通过此方法，我们可以提高客户分类的科学性，减少人为主观因素的干扰。然而，客户分类同时也是一种艺术，它要求企业在客户分类中必须从企业和客户客观实际出发进行合理分类。在具体的客户分类工作中企业必须应用科学方法，发挥创造性的艺术，做出科学和艺术的客户分类，从而服务于企业的发展目标。希望本文的研究工作对今后的客户分类管理研究能有所帮助。

[参　考　文　献]

[1]黄亦潇，邵培基，李菁菁.基于客户价值的客户分类方法研究[J].预测，2004，23(3)：31-35

[2]方开泰.实用多元统计分析[M].上海：华东师范大学出版社，1989

[3]余锦华，杨维权.多元统计分析与应用[M].广州：中山大学出版社，2005.2

[4]白奕.多指标综合评价的主成分分析模型及原理[J].陕西师范大学学报(自然科学版)，1998，26(2)：105-106

[5]宋艳，梁静国.基于模糊聚类的客户分类应用研究[J].物流科技，28(113)：26-28

[6]李玉民，李旭宏，毛海军，顾志康.主成分聚类分析在省域物流规划的应用[J].东南大学学报(自然科学版)，2004，34(7)：549-552

Researching on customers' classification in view of principal component-cluster analysis method

WANG Jian-min, WANG Chuan-xu

Abstract: The better foundation of customers' classification is the profit contribution degree of the customer to the business enterprise. In this paper, making use of the principal component-cluster analysis method, we research on the profit contribution degree and carry on the customers' classification by setting up the original data matrix of trade record.

Key words: Customers' Classification; Profit Contribution; Principal Component Analysis; Cluster Analysis