基于系统聚类的客户分析

李 斌[1] ,郭剑毅[2]
(1. 昆明理工大学 生物与化学工程学院 ,云南 昆明　650224 ;
2. 昆明理工大学 信息工程与自动化学院 ,云南 昆明　650093)

摘要: 采用最小方差法的系统聚类改进算法以减小异常数据对聚类过程的不利影响 ,通过对实际调查的样本数据进行聚类分析 ,挖掘和分析客户关系管理中客户群所存在的不同特征的组群 ,得到了直观的聚类过程和较合理的分组结果.

关键词: 数据挖掘 ;聚类分析 ;客户关系管理
中图分类号:TP311. 11　　**文献标识码**:A　　**文章编号**:1007 - 855X(2004)06 - 0066 - 04

Customer Analysis Based on Hierarchical Cluster

LI Bin[1] , GUO Jian-yi[2]
(1. Faculty of Biological and Chemical Engineering , Kunming University of Science and Technology , Kunming 650224 , China ;
2. Faculty of Information Engineering and Automation , Kunming University of Science and Technology , Kunming 650093 , China)

Abstract : In order to reduce the adverse effect on the cluster course of unusual data , an improved method of Hierarchica1 C1uster of Minimum Variance to analyze sample data of investigation is used , which can discover varied characters of different groups of customers in Customer Relationship Management. Cluster course and relatively rational grouping result are directly perceived through the senses.

Key words : data mining ; clustering ; customer relationship management

0 引 言

在“客户经济 ”时代 ,企业正在从以产品为中心向以客户为中心转变 ,强调的是以个性化、客户化的产品来满足不同区域市场在不同生命周期的需求[5].

面对错综复杂的客户信息和各种各样的消费习惯、消费意识和消费行为 ,采用聚类分析可以帮助市场人员发现客户群中所存在的不同特征的组群 ,并可以利用购买模式来描述这些具有不同特征的顾客组群 ,对客户行为聚类和分组可以帮助企业有策略地定义市场营销计划 ,选择营销渠道及策划广告以达到改善客户关系 ,并对将来的趋势和行为进行预测 ,支持企业决策[4].

1 聚类与系统聚类分析

1.1 聚类分析

聚类(clustering)是数据挖掘中的技术和手段之一 ,它是指将一个数据集划分为若干组(class)或类(cluster)的过程 ,并使得同一个组内的数据对象具有较高的相似度 ,而不同组内的数据对象则是不相似的[1]. 聚类分析是一种无指导学习方法和通过观察学习方法 ,面对一批样本 ,不知道它们的分类 ,甚至连分成几类也不知道 ,希望用某种方法把观测进行合理的分类 ,使得同一类的观测比较接近 ,不同类的观测相差较多.

1.2 系统聚类分析

系统聚类属层次聚类方法[1] ,是一种逐次合并类的方法 ,最后得到一个聚类的二叉树聚类图. 其方法

收稿日期:2003 - 12 - 01.
第一作者简介:李斌(1965～) ,男 ,硕士研究生. 主要研究方向:管理信息系统、数据挖掘. **E - mail** :kmlb @vip. sina. com.

是:对于 n 个观测,先计算其两两的距离得到一个距离矩阵,然后把离得最近的两个观测合并为一类,于是我们现在只剩了 $n-1$ 个类(每个单独的未合并的观测作为一个类). 计算这 $n-1$ 个类两两之间的距离,找到离得最近的两个类将其合并,就只剩下了 $n-2$ 个类 ……直到剩下两个类,把它们合并为一个类为止. 当然,真的合并成一个类就失去了聚类的意义,所以上面的聚类过程应该在某个类水平数(即未合并的类数) 停下来,最终的类就取这些未合并的类. 决定聚类个数是一个复杂的问题[2].

设观测个数为 n,变量个数为 v,G 为在某一聚类水平上的类的个数,x_i 为第 i 个观测,C_k 是当前(水平 G) 的第 K 类,N_k 为 C_k 中的观测个数,$\overline{X}$ 为均值向量,$\overline{X}_k$ 为类 C_k 中的均值向量(中心),$\|x\|$ 为欧氏长度,$T^2=\sum_{i=1}^{n}\|x_i-\overline{X}\|^2$ 为总离差平方和,$W_k=\sum_{i\in C_k}\|x_i-\overline{X}_k\|^2$ 为类 C_k 的类内离差平方和,$P_G=\sum W_J$ 为聚类水平 G 对应的各类的类内离差平方和的总和. 假设某一步聚类把类 C_k 和类 G_L 合并为下一水平的类 C_M,则定义 $B_{KL}=W_M-W_K-W_L$ 为合并导致的类内离差平方和的增量. 用 $d(x,y)$ 代表两个观测之间的距离或非相似性测度,D_{KL} 为第 G 水平的类 C_K 和类 C_L 之间的距离或非相似性测度. 进行谱系聚类时,类间距离可以直接计算,也可以从上一聚类水平的距离递推得到,根据类间距离的计算方法的不同,有多种不同的聚类方法[7].

2 一种带约束的最小方差系统聚类法

算法介绍:对于较小容量的数据库或数据样本,由于其代表的局限性,有一些数据具有奇异性,当对含有奇异数据进行聚类时,可能干扰判别类数的结果,从而影响聚类效果,误导聚类分析. 如果对样本数据先进行数据处理,以减小奇异数据的影响或者降低奇异数据的干扰,可以增强聚类过程的有效性和说服力.

方法是先对客户数据集进行线性变换,把结果放到另一个数据库中,再利用最小方差法的谱系聚类对新数据集进行聚类分析,最后得到该组客户的分类情况.

具体计算公式如下:

类内离差平方和为:

$$W_i=\sum_{l=1}^{n_i}(x_l-\overline{x}_l)'(x_l-\overline{x}_l) \tag{1}$$

类间离差平方和为:

$$P_k=\sum_{i=1}^{k}W_i=\sum_{i=1}^{k}\sum_{l=1}^{n_i}(x_l-\overline{x}_l)'(x_l-\overline{x}_l) \tag{2}$$

当 k 固定时,应选择使 P_K 达到最小的分类.

递推公式为:

$$D_{JM}=((N_J+N_K)D_{JK}+(N_J+N_L)D_{JL}-N_JD_{KL})/(N_J+N_M) \tag{3}$$

利用最小方差法并类时总是使得并类导致的类内离差平方和增量最小.

算法实现步骤为:

(1) 对数据进行预处理,既对数据集中关心的属性作线性变换处理,把结果放到一个新的数据库中;

(2) 采用最小方差法对新数据集的数据进行分析;

(3) 进行立方群标准、伪 F、半偏 R^2、R 和伪 t^2 统计等检验指标计算,初步确定类水平数;

(4) 进行主因素分析;

(5) 确定样本分类数.

3 系统聚类分析的实现

在客户关系管理中,聚类分析可以用于客户分析、对手分析、合作伙伴分析、需求分析、市场营销分析、销售分析、价格分析等. 这里根据客户的购买行为、购买方式、购买频率、购买金额等因素对客户群进行划

分.

3.1 数据采集

数据来源于调查结果,调查地点一个是开远市的"解放军化肥厂"和"红河州化肥厂",另一个是安宁市的"云南磷肥工业基地",分别发出调查表200份,各收回120分和60份,调查时间为2004年元月.

3.2 调查结果

调查结果涉及23个变量,部分数据见表1.

表1 调查结果

Tab.1 Data of investigation

编号	性别	婚姻	年龄	工龄	上月总收入/元	生活费比例/%	副食比例/%	通讯费比例/%	总支出比例/%
1 001	Male	Single	26	7	1 000	40	10	20	85
1 002	Male	Married	33	13	900	76.67	2.22	4.44	105.56
1 003	Female	Single	28	8	1 060	66.04	18.87	7.55	92.45

3.3 聚类过程

采用SAS编程语言,首先把数据库中的"安宁市调查表"数据放入数据集SASUSER.ANNING中,考察四个变量:生活费比例(LivingComparison)、副食比例(foodComparison)、通讯费比例(Communication-Comparison)、总支出比例(TotalComparison).

聚类过程部分代码为:

```
proc aceclus data = sasuser.anning
out = ace p = .03 noprint;
   Communication Comparison
Total Comparison;
run;

proc cluster data = ace outtree = tree
   method = ward
var LivingComparison
food Comparison
 ccc pseudo ;
var can1 can2 can3;
id CustomerId;
run;
```

计算结果见表2.

表2 协方差特征值表

Tab.2 Eigenvalues of the covariance matrix

因子	特征值	差值	贡献率	累计贡献率
1	66.377 655 6	61.155 014 3	0.897 5	0.897 5
2	5.222 641 3	2.865 552 0	0.070 6	0.968 1
3	2.357 089 3	2.357 089 3	0.031 9	1.000 0

总样品均方根的标准偏差 = 4.965 125,不同观测之间的均方根距离 = 12.162 02.

表2贡献率是指特征值占总方差的百分比,从表2结果可见,第一个特征值66.37,贡献率达到为89.75%,而且前三个成分对数据的解释能力达到了100%.根据比例标准可选取两个主因子.

聚类过程见表3,由于篇幅所限只选类水平10到类水平1的聚类过程.

调查结果涉及23个变量,部分数据见表1.

表3,NCL代表聚类水平,ClusterJined代表本次聚类将哪两类合并在一起,FREQ表示本次合并的类有多少个观测,SPRSQ是半偏R^2,RSQ是R^2,ERSQ是零

表3 聚类过程

Tab.3 Cluster History

NCL	Clusters	Joined	FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2
10	CL21	CL18	9	0.004 4	0.967	0.943	5.36	153	9.7
9	CL31	CL36	4	0.005 4	0.962	0.936	5.04	150	28.6
8	CL14	CL17	4	0.005 7	0.956	0.928	4.91	152	3.3
7	CL11	CL10	35	0.009 6	0.946	0.918	3.52	147	13.8
6	CL7	CL19	39	0.010 7	0.936	0.905	3.32	148	11.5
5	CL6	CL16	47	0.012 8	0.923	0.886	3.37	155	11.6
4	CL9	CL8	8	0.029 9	0.893	0.859	2.47	147	12.0
3	CL4	1 012	9	0.052 3	0.840	0.812	1.13	142	8.1
2	CL5	L3	56	0.173 5	0.667	0.683	- 0.41	110	58.7
1	CL2	1 015	57	0.667 0	0.000	0.000	0.00	0.0	110

假设期望值,CCC 为立方群聚标准,PSF 是伪 F 统计,PST2 是伪 t^2 统计[6].

在聚类分析中必须确定一个合理的分类数目,这个数目可以由输出的 CCC、PSF、PST2 等这些量决定,一般在 CCC 和 PSF 出现峰值所对应的分类数较合适,在 PST2 出现峰值的前一行所对应的分类较合适.

因为我们事先并不知道数据的实际分类情况,所以我们必须找到一个合理的分类个数. 为此,考察 CCC、伪 F 和伪 t^2 统计量.

根据上面产生的 TREE 数据集,绘制各统计量的图形见图 1～3 所示:

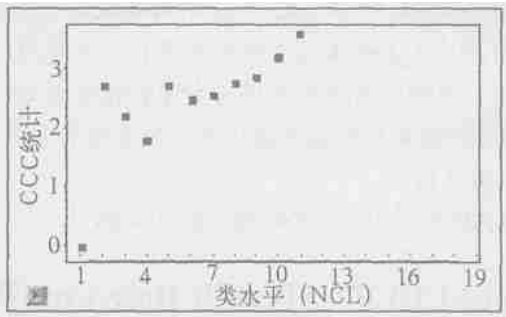

图 1　立方群标准统计散点图

Fig.1　Scatter plot of CCC

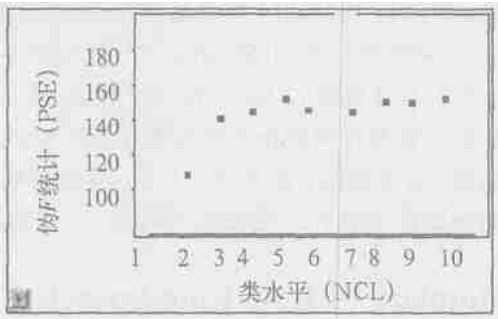

图 2　伪 F 统计散点图

Fig.2　Scatter plot of PSF

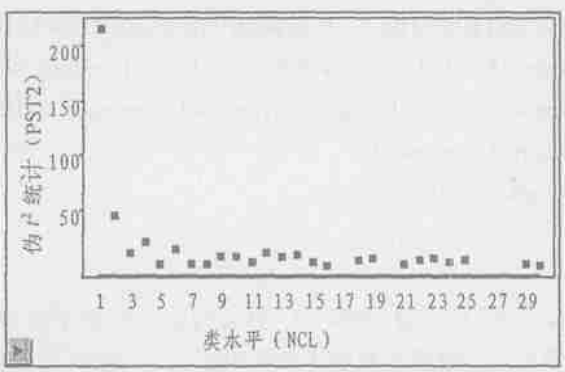

图 3　伪 t^2 散点图

Fig.3　Scatter plot of PST2

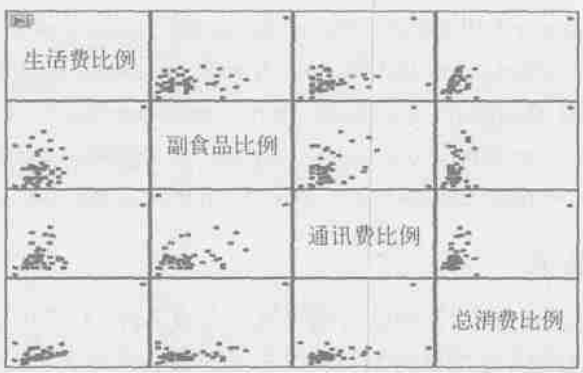

图 4 散点图矩阵

Fig.4　Scatter plot of matrix

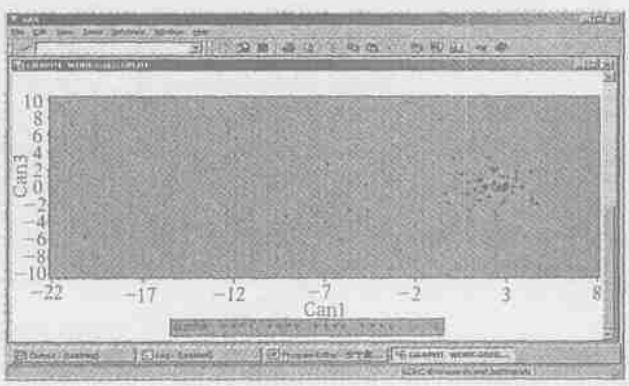

图 5　主分量的散点图

Fig.5　Scatter plot of principal factors

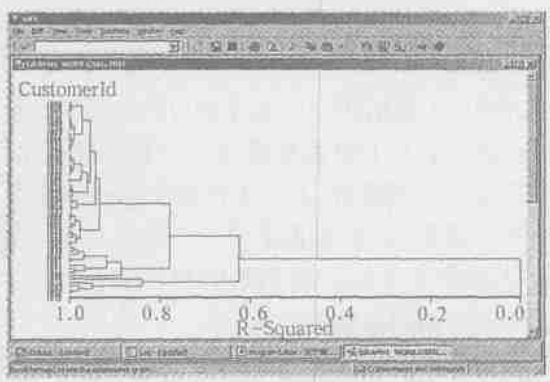

图 6　树状图

Fig.6　Tree of cluster

分析这三个统计量,可以发现,在 CCC 散点图中,局部最大值在 2 和 5 处,建议取 2 类或 5 类;伪 F 散点图中的局部最大值在 5 处,建议取 5 类;再看伪 t^2,建议 3 类或 5 类(局部最大值处是不应合并的,即局部最大值处的类数加 1);半偏 R^2 统计建议取 3 类. 由这些指标可以判断比较一致的是 3 类,其次是 5 类. 为了最终确定分类数,下面来比较一下四个变量的散点图矩阵,见图 4.　(下转第 74 页)

法.很明显这种算法最大的缺点就是延迟太大,一帧的延迟是60 ms(包括编码与解码),主要原因是帧长太长,子帧7.5 ms比 G.729 的子帧5 ms还长.因此这种编码算法更加适合于低速率语音传输系统中.

参考文献:

[1] 成德源.8 kb/s低复杂度代数码激励线性预测编码[J].电路与系统学报,1997,2(4):44～48.

[2] Soong F K, Juang B H. Line spectrum pair(LSP) and speech data compression[J]. Proc. IEEE ICASSP,1984.

[3] Terrestrial Trunked Radio (TETRA);Speech codec for full - rate traffic channel;Part 2: TETRA codec, ETSI ETS 300 395 - 2 ed. (2):1998～02.

(上接第69页)

比较矩阵中几个结果,相对集中有 LivingComparison 和 TotalComparison 的散点图,foodComparison 与 TotalComparison 的散点图,通过以上分析,确定 LivingComparison 和 TotalComparison 为主分量.

下面分别用 gplot 和 tree 方法[3][6]作以上两个主分量的散点图和树状图,见图 5 和图 6.

图 5 中用不同颜色直观地显示不同的类,可以看出划分为 3 类比较合理,从而证实了前面分析结果的有效性.图 6 中从右端开始一类分为两类,到了最左端完全是每一个观察一类了,而且标出了观察的客户编号,由于客户较多不宜纵向展开,所以显得很乱,但基本可以看出分为 3 类比较合理.

综合以上讨论和分析可以得到结果:根据样本数据将考察客户群分为三类.

4 结论

利用数据挖掘技术中的聚类分析,可以把考察的客户群进行分类,这样对不同的客户进行定位,以便针对不同的客户采取不同的营销策略,以及进行不同的服务内容.在这里利用类间最小方差法的改进系统聚类分析方法对实际调查样本数据进行聚类分析,不但可以直观地看到聚类过程,而且根据技术分析可以得到聚类结果,从主分量散点图和树状分类图可以清楚地判别分组情况是比较合理的,从而证实了模型的正确性和有效性.

参考文献:

[1] 朱明. 数据挖掘[M]. 合肥:中国科学技术大学出版社,2002.129～162.

[2] 张维明. 数据仓库原理与应用[M]. 北京:电子工业出版社,2002.182～206.

[3] 朱扬勇,左子叶. 数据挖掘实践[M]. 北京:机械工业出版社,2003.159～201.

[4] 堂林燕. 数据仓库和数据挖掘技术在 ERP 中的应用[J]. 上海:计算机工程,2002,28(9):10～18.

[5] 杨东龙. 客户关系管理[M]. 北京:中国经济出版社,2002.106～138.

[6] 樊欣. SAS 经济统计[M]. 北京:北京希望电子出版社,2003.259～270.