基于客户流失分析的聚类分析模型的构建

曾 瑞[1] 胡建华[2] 高 敏[3]

(云南师范大学计算机与信息学院计算机科学系 昆明 650092)[1]

(昆明理工大学计算机与自动化控制学院计算机科学系 昆明 650093)[2] (昆明市电信局 昆明 650033)[3]

摘 要 本文针对电信公司的客户流失主题,采用了聚类方法对特征属性进行分类。鉴于聚类算法的无监督性,结合因子方差分析方法对聚类进行监督,分析特征属性与目标变量的相关性,利用相关属性与目标变量共同参与聚类算法构建聚类模型,获取流失客户的特征。

关键词 数据挖掘,聚类,因子方差分析法

A Clustering Mode of Missed Customer

ZENG Rui[1] HU Jian-Hua[2] GAO Ming[3]

(Yunnan Normal University, Kunming 650092)[1] (Kunming University of Science, Kunming 650093)[2] (Kunming Telecom, Kunming 650033)[3]

Abstract Aimed at the customer missing in telecom, clustering is used to classify the characterized vector in this article, and for the non- supervision of clustering, factor variance analysis is used to analyze the relevance between characterized vector and target variable, and then the relevant characterized vector and target variable are used together to establish the clustering model, and finally the features of missed customer were acquired.

Keywords Data mining, Clustering analysis, Factor variance analysis

市场竞争日趋激烈,应用数据仓库技术进行数据分析,建立营销机制,已成为经营决策者的共识。如何从大量的、不完全的、有噪声的、模糊的、随机的数据中,提取隐含在其中的、潜在的有用信息[1],正是数据挖掘的过程。通过数据挖掘技术能有效地细分客户群体,使经营者能及时地制定相应的营销政策,用最小的花费得到最好的回报。如何选取数据挖掘方法进行有效的数据挖掘已成为热点问题。本文拟针对某电信公司客户流失主题的分析,尝试利用统计分析方法中的因子分析法对聚类算法进行监督,分析各特征属性与流失主题是否相关,去除样本中无关属性,建立聚类模型,对客户进行有效的分类,为营销策略的建立提供有力的依据。

1 建立目标变量

目标变量是能反映分析主题的变量。目标变量不一定是已存在的属性,若不存在则需构建。只有目录变量参与到聚类算法中,建立的模型才能反映分析主题。

客户流失分析中,选取流失前12个月的历史消费信息作为样本数据。在样本数据中包含流失的客户的基本属性(如流失时间、所在部门、所属产品、是否VIP客户、营销属性、统计属性、属地、费率类型等),及客户在流失前一年(自流失的月份向前推12个月)的消费情况。但这些属性是否能单独地刻画客户的流失趋势?回答是不一定的,OLAP能进行相应的分析。例如,分析流失客户在流失前的月消费额如图1。

由图1可以看出,流失用户在流失前的消费额并不是逐月减少的,把每个用户的月消费额作为目标维变量是没有意义的。类似地,利用OLAP可以分析出单独的用户基本属性不能描述客户流失的趋势,但基本属性的组合——通话费与其总消费额的比值则可以反映流失客户的特征。因此目标维变量确定为这个比值,在本文中简称为消费比。在样本数据中添加消费比属性,完善样本集。

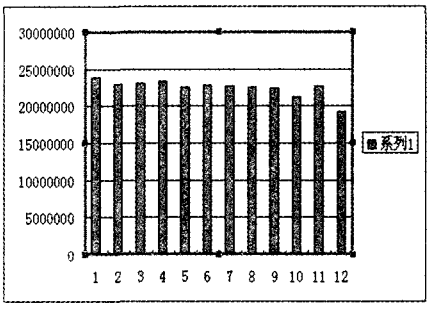

图1 月消费额趋势

2 利用因子方差分析法分析与目标变量相关的属性

聚类和分类之间的不同就在于:分类问题中在分类前已经知道了训练例的分类属性,而聚类算法是无监督的,需要在训练例中寻找分类属性。过多的属性参与反而会干扰聚类算法寻找分类属性,得出的数据挖掘模型在分出的类别上将不具有显著的差异特征,不能满足实际要求。而针对主要特征分量进行的聚类才能得到合理的分析模型。

本文选用了统计分析方法中的因子方差分析法对聚类算法进行监督,分析各特征分量与目标变量是否相关。

因子方差分析问题就是在方差相等情况下对多个正态均值是否彼此相等的一个假设检验问题,所涉及的一对假设如下:

$H_0:\mu_1=\mu_2=\cdots\mu_r$, H_1:诸 μ_i 不全相等[2]。

在原假设 H_0 成立下,两个均方和之比服从 F 分布,即

$$F=\frac{MS_A}{MS_e}\sim F(r-1,n-r)$$[2]

曾 瑞 讲师,学士,主要研究方向:数据仓库、多维数据分析;胡建华 副教授,主要研究方向:数据仓库、数据挖掘;高 敏 助工,硕士,主要研究方向:数据仓库、多维数据分析。

其拒绝域应为：$W=\{F>c\}$，对给定的显著性水平 $\alpha(\alpha=0.05)$，c 由 F 分布的 $1-\alpha$ 分位数 $F_{1-\alpha}(r-1,n-r)$ 确定。

当各正态均值有显著差异时，假设发生 $F>F_{1-\alpha}(r-1,n-r)$ 的概率为 p，则称 p 为显著性概率[3]。若 p 值小于 α，则拒绝零假设，各组均值有显著差异，即因子与考察指标有关。反之，若 p 值大于 α，则不能拒绝零假设，各组均值无显著差异，即因子与考察指标无关。

将客户流失样本中的各基本属性与目标变量输入程序，即可得出图 2 的分析结果。

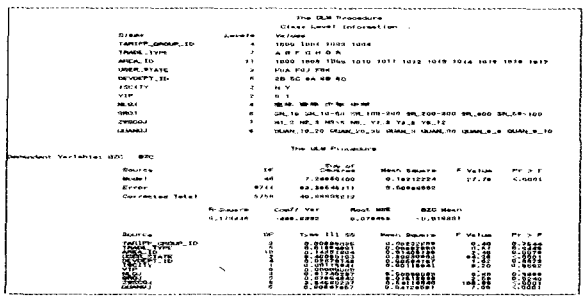

图 2　因子分析：目标维变量_各特征分量

由图 2 可以得出，与目标变量无关的属性有：TRAIFF_GROUP_ID(费率类型)、TRADE_TYPE(行业)、DEVDEPT_ID(营销、统计属性)、ISCITY(地处)、VIP(VIP 客户)、NLQJ(年龄区间)；与目标变量有关的属性有：AREA_ID(地域)、USER_STATE(用户状态)、SRQJ(消费额区间)、ZWSCQJ(在网时长区间)、QUANQJ(话务量区间)。

3　客户流失聚类模型的建立

聚类分析是数据挖掘中的一种主要技术，是把一组个体按照相似性分成若干类别，使得属于同一类别的个体之间的距离尽可能地小，而不同类别上的个体间的距离尽可能地大。

聚类分析主要分为层次聚类和迭代的平方误差分区聚类。层次方法按群组的嵌套顺序组织数据，以树状图或树形结构来表示。平方误差分区算法试图得到一个使类内分散最小而类间分散最大的分区[3]。

在客户流失分析中采用的就是迭代的平方误差分区聚类。设 m 是样本参与聚类的属性个数，n 是样本的个数，S 是由用户预先设定的分类数目，聚类分析问题可描述为：给定 m 维空间 R^m 中的 n 个向量，把每个向量归属到 S 聚类中的某一个，使得每个向量与其聚类中心的“距离”最小[4]。

根据上述因子方差分析的结果，剔除样本数据中与分析主题无关的属性，保留相关属性参与聚类算法，得出聚类模型，如图 3。

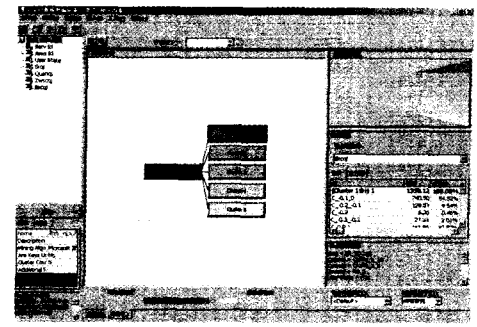

图 3　针对相关特征分量的客户流失聚类模型

由图 3 可以看出，聚类时选取了相关属性，把集合聚合为 5 类。例如，第一类节点的数据，根据每个字段代表的意义，可以解释为：消费额区间为 50～200 元，话务量区间为 3～20 个小时，在网时长区间为 1～5 年，用户状态为正常流失，地域为营销中心。从特性数据区中，可以看出这类人群的消费比在－0.1～0 区间时的流失的可能性为 54.52%，消费比在－0.1～－0.2 时的流失的可能性为 9.54%，……。

这类人群已经有了较为显著的特征：消费额、话务量、在网时长都处于中等水平，且为主动流失的，而且限定了地域范围。基本达到了客户细分的功能，为营销策略的构建提供了依据。

总结　本文针对电信客户流失主题，采用了聚类方法对特征属性进行分类，鉴于聚类算法的无监督性，结合因子方差分析方法对聚类进行监督，构建的聚类模型，基本达到了细分客户的目的，能有效地反映流失客户不同群体的共同特征。从实践中证明，用该方法构建聚类模型是可行的方案。

参 考 文 献

1　邵峰晶，于忠清. 数据挖掘原理与算法. 中国水利水电出版社，2004. 2～15

2　高惠璇. 实用统计方法与 SAS 系统. 北京大学出版社，2001. 20～30

3　Chen M S, Han J, Yu P S. Data mining: An overview from database perspective. IEEE Transac-tions on Knowledge and Data Eng, December 1996. 866～883

4　Fayyad U. Data Mining and Knowledge Discovery. Making Sense Out of Data. IEEE Expert Intelligent Systems and their Applications, 1996

(上接第 181 页)

术。该算法具有很强的实用性，因为网络的拓扑结构采用星型结构，这样网络结构更加适合现实系统的应用现状。另外，从网络的建设的成本以及网络的管理和维护方面比 FDM 系统采用的网状结构相比较，具有成本低及易于管理的优点。

在大规模的分布式数据库中进行关联规则的挖掘是数据库技术与并行处理技术两者相结合的一种方法。由于数据规模的进一步扩大，分布式关联规则的网络结构将会变化成为树形结构。本算法经改变，也可以适合树形结构的网络拓扑结构。

参 考 文 献

1　Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of data, 1993. 207～216

2　宋宝莉，覃征. 分布式环境下关联规则的安全挖掘算法[J]. 计算机工程，2006，21

3　陈涛，石伟胜，陈启买. 关联规则的并行挖掘算法研究[J]. 现代计算机，2006(7)

4　Bayardo R, Agrawal R. Mining the most interesting rules. In: Proc. of the ACM SIGKDD Conf on knowledge Discovery and Data mining, San Diego, CA USA, 1999. 145～154

5　Cheung D W, han J, Ng V, et al. A fast distributed algorithm for mining association rules. In: Proc. 1996 Int Conf. Parallel and Distributed Information Systems. Miami Beach, Florida, Dec. 1996. 31～44

6　陈耿，倪巍伟，等. 基于分布数据库的快速关联规则挖掘算法[J]. 计算机工程与应用，2006(4)

7　Schuster A, Wolff R. Communication-efficient Distributed Mining of Association Rules. In: Proc. 2001 ACM SIGMOD Int Conf. Santa Barbara, California, May 2001. 473～484