

聚类分析在燃气用户细分中的应用

西安市天然气总公司(710082) 于鸿潮 梁忠伟

1 前言

燃气公司有为数众多的各类用户,在经营过程中会接触大量的信息和数据,如何能成功地收集分析理解信息,并根据这些信息有效的决策,是燃气公司信息化过程中的关键任务。随着计算机在企业管理的广泛应用,“数据泛滥,信息缺乏”这一现象变得比较明显。燃气企业通过科学的方法从大量的历史数据中找出对企业决策有帮助的信息,从企业存储的大量的客户历史记录挖掘中,不断改善与客户的关系,识别对企业利润贡献大的用户,挖掘客户终身价值,可以使企业得到更快更长远的发展。

传统上,燃气公司习惯于将众多燃气用户按预先设定好的分类划分,如普通居民、锅炉、直燃机、餐饮、工业窑炉(生产线)、加气站等。并根据用户的分类确定服务政策和办法,分配服务资源。但这种燃气用户细分方法的缺点是显而易见的,它是机械和静态的,没有考虑到用户的差异性和发展变化,更主要的是,这种划分是基于技术层面的,没有考虑到不同用户对于公司的终身价值问题。数据挖掘技术中的聚类分析方法为科学、合理的进行燃气用户细分提供了有力的工具。

2 聚类分析简介

面对海量的资料和数据,首要的任务是将它合理归类,否则,在许多情况下,问题不是模式没有被发现,就是模式太多了。聚类分析就是将数据合理归类的一种方法。聚类就是将数据对象分组成为多个类或簇,在同一个簇中对象之间具有较高的相似度。聚类不同于分类,在分类模块中,对于目标数据库中哪些类是知道的,要做的就是将每条记录分别属于哪一类标记出来;而聚类所要的划分类是未知的,也就是在对目标数据库到底有多少类预先不知道的情况下,希望将所有的记录组成不同的类。这种技术对发现基于相似特征(如人口统计信息、财政信息或购买行为)的客户细分非常有用。

3 用户的数学描述和细分标准

为了建立科学的燃气用户细分模型,对用户进行恰当的数学描述是很重要的。为了说明用户聚类分析的过程,在考虑了用户对燃气公司的价值、对供气均衡性的影响、维修服务的需求等因素,本文作者确立以下6个指标作为描述用户的变量:

多渠道的融资方式,各方位的资金注入,构成企业充足的初始投资,使得企业拥有完善的城市管网、场站设备等基础设施,这些充分显现了公司的资金优势。但是,西安市天然气公司虽拥有相当的优势,可产品市场单一,盈利能力较弱。虽然基本完成了西安市“煤改气”的目标,企业仍有很大的发展潜力,完善的基础设施未能充分发挥作用。随着汽车产业的高速发展,带来了能源短缺,环境污染两大世界问题,而

天然气汽车的应用恰好从根本上解决了这些问题,目前加气站的用气量在企业内部所占份额较大,并且发展前景乐观。石油资源的短缺和日益严重的大气污染所面临的困境,为天然气汽车加气站的建设提供了广阔的市场机遇和远大的发展空间。西安市天然气公司应充分利用难得的市场机遇,根据自身现状使企业各种优势得以完全发挥,在未来越演越烈的竞争中立于不败之地。

—— 年平均用气量(AVY):用户自投运来总消费气量/运行时间,主要反映用户的用气规模,以便确定该用户对公司利润贡献的大小。

—— 高峰用气占全年的比例:北方城市采暖期的4个月消费气量/全年消费气量,主要反映用户用气的均匀性。有些用户虽然年平均用气量(AVY)较大,但主要集中在采暖期的个别月份,高峰时对管网的运行带来很大压力,低峰时对公司没有丝毫贡献,这类用户在细分模型中要与那些均匀性用户区分开。

—— 年平均维护工时:年维修,抢修,抄表,安全检查等消耗的工时,主要反映维护成本。不同用户对于维修、维护、检查等的服务需求是不同的,设立这一指标也是为了反应用户对燃气公司实际利润的贡献。

—— 实际用气量与最大供气能力的比值:实际用气量/该用户设计最大供气能力,主要反映该用户的发展潜力。公司对用户的供气能力由于技术原因是离散的,比如选择D200管径的用户的最大供气能力是选择D150用户1.76倍。为了保证用户正常用气,用户接入设备的最大供气能力都要不同程度地大于设计用气能力,与实际用气量的差异就更大。这一指标可以反应用户在不进行额外设备投资情况下,用气增长的潜力。

—— 天然气消费量占全部清洁能源消费量的比例:天然气消费量/电力、柴油、天然气等全部清洁能源消费量,主要反映替代品对该用户的竞争程度和用户的增长潜力。很多用户都使用多种清洁能源,比较典型的是用燃气取暖,用电带动空调。如果公司对这类用户针对性地开展营销工作,对扩展燃气应用领域,平衡季节不均匀性都有很大好处。

—— 年用气量平均增长率:主要反映用户消费的增长情况。很多用户按照设计能力安装完天然气设备后用气并不是一成不变的,比如:某工业用户由于产品滞销,产能逐年下降,年用气量平均增长率也会呈现负数,这一指标就是反应这种情况。

以下选择某燃气公司接入管径同为D150的10个工商业用户进行聚类分析,对基于聚类分析的用户细分模型作出说明,具体见表1。

4 用户相异性的度量

4.1 数据的预处理

在进行聚类分析之前,需要对搜集到的数据进行预处理。目前,在聚类分析中使用的数据类型大多为下面两种:

- (1) 数据矩阵
- (2) 相异度矩阵

表1 D150工商业用户表

序号	用户名	年均气量	高峰比例	维护工时	能力	替代品	增长率
1	北二环加气站	55	30	2	75	75	10
2	省高院	1.5	95	10	30	30	-4
3	双佳电磁厂	12	80	15	60	50	5
4	省旅游汽车公司	25	85	15	65	55	14
5	嘉汇公司锅炉房	7	95	5	60	40	3
6	旅游局	12	95	18	90	95	1
7	荣和陶瓷	5	70	10	40	90	-8
8	省医药工业公司	8	80	1	31	80	6
9	省机械施工公司	47	30	4	70	70	12
10	205研究所	9	30	13	28	13	16

说明:表1中,年平均用气量的单位是万标准立方米;高峰用气比例=(11月15-3月15日用气量)/(全年用气量)×100%;能力=上年实际用气量/设计最大年用气量×100%;替代品=用户年缴纳天然气费/(天然气费+燃油费+电费+液化石油气费)×100%。

4.2 度量数据的标准化

由于度量单位的选取对于聚类结果有较大的影响,例如,将长度单位由米改为厘米,重量单位由吨改为公斤都将直接影响聚类分析的结果,为了避免出现这种情况,必须将数据进行标准化处理。

常用的标准化方法是减去其均值再除以其标准差,即:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad S_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

原始数据标准化值为:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$$

经过标准化处理的用户数据表如表 2 所示。

5 聚类分析算法的实现和用户细分模型的建立

聚类分析的算法较多,主要有以下几类:层次的方法、划分的方法、基于密度的方法、基于网格的方法和基于模型的方法。算法的选择主要取决于数据的类型、聚类的目的和应用。根据燃气用户的性质和用户细分的目的,本文作者选择了基于模型的方法进行聚类分析计算,所谓基于模型的方法即为每一类假定了一个模型,然后寻找能很好满足这个模型的数据集。

为了更好对这一方法加以说明,在 10 个用户中我们选择了 4 个比较典型的用户作为聚类模型,这 4 个用户有鲜明的特点,燃气公司应该采取有针对性的服务和营销措施,通过聚类分析将与这些用户

表 2 经过标准化处理的用户数据表

序号	用户名	年均气量	高峰比例	维护工时	能力	替代品	增长率
1	北二环加气站	1.680	-0.783	-0.808	1.355	0.494	0.198
2	省高院	-0.868	1.399	0.410	-0.739	-1.442	-1.109
3	双佳电磁厂	-0.368	0.896	1.171	0.657	-0.581	-0.269
4	省旅游汽车公司	0.251	1.064	1.171	0.890	-0.366	0.572
5	嘉汇公司锅炉房	-0.606	1.399	-0.351	0.657	-1.012	-0.456
6	旅游局	-0.368	1.399	1.628	2.053	1.355	-0.642
7	荣和陶瓷	-0.701	0.560	0.410	-0.273	1.140	-1.483
8	省医药工业公司	-0.558	0.896	-0.960	-0.692	0.710	-0.175
9	省机械化施工公司	1.299	-0.783	-0.503	1.122	0.279	0.385
10	205 研究所	-0.511	-0.783	0.867	-0.832	-2.174	0.759

表 3 4 个典型用户作为聚类模型的比较表

序号	用户名	年均气量	高峰比例	维护工时	能力	替代品	增长率	特点
1	西变三菱	85	35	8	70	60	4	利润贡献高、季节均衡、维护费用低
2	权瑞科技	45	99	20	95	95	0	利润较高, 季节不均衡、潜力小
3	万年饭店	12	30	30	30	40	30	利润一般、有潜力、增长快
4	荣和陶瓷	5	70	10	40	90	-8	利润低、增长慢或负增长

表 4 聚类分析计算结果表

用户名	To 类型 1	To 类型 2	To 类型 3	To 类型 4	Minimum	分类
北二环加气站	3.688	15.585	32.509	14.850	3.688	类型 1
省高院	25.649	23.737	24.555	7.757	7.757	类型 4
双佳电磁厂	15.912	10.076	15.619	6.107	6.107	类型 4
省旅游汽车公司	13.090	8.329	14.311	9.583	8.329	类型 2
嘉汇公司锅炉房	19.028	16.841	27.616	7.845	7.845	类型 4
旅游局	21.673	2.643	28.832	8.463	2.643	类型 2
荣和陶瓷	20.856	14.046	28.630	0.000	0.000	类型 4
省医药工业公司	20.930	21.476	30.342	4.081	4.081	类型 4
省机械化施工公司	4.417	15.078	26.406	12.816	4.417	类型 1
205 研究所	22.870	33.846	9.789	18.366	9.789	类型 3

类似的用户划分出来,对实际工作有重要的指导意义。具体见表 3。

进行聚类分析决定某一用户 n 的分类时要进行

如下计算:

分别算出用户 n 标准化的各变量值与这 4 类用户标准化的各变量值的平方差的和 Z

$$\begin{aligned}
 Z_{n24} &= (X_{na}-X_{24a})^2 + (X_{nb}-X_{24b})^2 + (X_{nc}-X_{24c})^2 + (X_{nd}-X_{24d})^2 \\
 &\quad + (X_{ne}-X_{24e})^2 + (X_{nf}-X_{24f})^2 \\
 Z_{n34} &= (X_{na}-X_{34a})^2 + (X_{nb}-X_{34b})^2 + (X_{nc}-X_{34c})^2 + (X_{nd}-X_{34d})^2 \\
 &\quad + (X_{ne}-X_{34e})^2 + (X_{nf}-X_{34f})^2 \\
 Z_{n25} &= (X_{na}-X_{25a})^2 + (X_{nb}-X_{25b})^2 + (X_{nc}-X_{25c})^2 + (X_{nd}-X_{25d})^2 \\
 &\quad + (X_{ne}-X_{25e})^2 + (X_{nf}-X_{25f})^2 \\
 Z_{n7} &= (X_{na}-X_{7a})^2 + (X_{nb}-X_{7b})^2 + (X_{nc}-X_{7c})^2 + (X_{nd}-X_{7d})^2 + (X_{ne}-X_{7e})^2 \\
 &\quad + (X_{nf}-X_{7f})^2
 \end{aligned}$$

然后在 Z_{n7} — Z_{n34} 中取最小值,最小值对应的典型用户就是该用户的分类。

聚类分析计算结果如表 4 所示。

6 小结

在实际使用中,可以根据需要建立对用户更详细的数学描述,还可以选择更多的用户模型,使用户细分更加准确。用户数据要能通过公司信息系统动态更新,并注意根据最新的细分结果,灵活地调整用户服务政策,合理地分配服务资源。

参考文献

- 1 刘书香,卢才武.数据挖掘中的客户聚类分析及其算法实现,《信息技术》[J].2004 年 1 月,第 28 卷第 1 期,PP.4-10
- 2 A.K.Jain,M.N Murty and P.J.Flynn.Data clustering:A Survey.[J].ACM Comput.Surv.31:PP.264-323,1999
- 3 赵绍忠,戴稳胜.数据挖掘中的聚类分析,《统计与信息论坛》[J].2002 年 5 月,第 17 卷总第 53 期,PP.4-10
- 4 Herb Edelstein, 浅说数据挖掘,《计算机系统应用》[J].1998 年第四期,P56-57