# 基于 K-means 聚类与决策树的有线电视交互服务订制预测

杨怀珍,李玲华
(桂林电子科技大学 管理学院,广西 桂林 541004)

**摘 要**: 建立了一种基于聚类分析与决策树分析相结合的服务订制预测模型,阐述了聚类分析 K-means 算法、决策树算法 C5.0 算法原理、建模流程的设计,将模型应用于某地区用户对有线电视交互服务的订制意愿预测,最终确定高响应率客户群。实验证明,该模型相对于仅通过决策树进行预测能更大程度地提高分类精度,并能更有效地识别出高响应率客户群。

**关键词**: K-means 聚类;决策树;C5.0;有线电视交互服务
**中图分类号**:TP84+ **文献标识码**: B

# Order prediction of interactive service of CATV based on K-means cluster and decision tree

YANG Huai Zhen, LI Ling Hua
(School of Management, Guilin University of Electronic Technology, Guilin 541004, China)

**Abstract**: This paper builds a forecasting model of service marketing which is based on cluster analysis combining with decision tree. It depicts K-means algorithm, C5.0 algorithm of decision tree and design of building model, and applies the model on predicting whether a region users will accept the interactive service of CATV, by this way it finds the users group which shows the highest response rate. The result shows that the model finds the users group of high response rate easier than predicting only by decision tree, it also improves the classification accuracy on a greater extent.

**Key words**: K-means clustering; decision tree; C5.0; interactive services of CATV

决策树技术早已被证明是利用计算机模仿人类决策的有效方法。20 世纪 80 年代,它是构建人工智能系统的主要方法之一。20 世纪 90 年代初,这一技术随着人工智能遭遇低潮而逐渐不为人所注意。然而,20 世纪 90 年代后期,随着数据挖掘技术的兴起,决策树又重新引起了人们的重视。在决策树研究应用中,有学者侧重于研究产生决策树的训练和检验数据的大小及特性与决策树特性之间的关系,即着重于产生决策树的数据。一些专家认为,在产生决策树前尽量减少训练数据量比在决策树产生后再简化决策树更能够提高决策树的性能,实际上,这就是经常提起的数据预处理技术,与决策树修剪技术相对应,也称为数据减少技术[1-4]。基于原始数据的缺陷及聚类的特性,将聚类应用于分类预测前的数据预处理中,可以删除冗余、相似、噪声数据,从而减少训练数据并提高训练数据的质量,进而改进单个决策树的性能。

## 1 相关理论及研究方法

### 1.1 决策树性能与训练数据间的关系

Oates 等人[5]研究了训练集的大小与决策树复杂性之间的关系,其研究表明训练数据的增加经常会造成决策树大小的线性增加,但这种增加并没有带来决策树分类准确性的提高。Sebban 等人[6]研究了训练集的质量和大小对决策树的影响,包括训练出的决策树模型的复杂性与泛化精度之间的关系,并从理论上论证了可以在不影响分类精度的前提下通过减少训练数据来减小决策树。Sebban 等人提出了使用原型选择算法来减少原始数据而提高后剪枝决策树的性能。当训练集的大小随机减小时,决策树的大小也随之减小,并且不影响分类精度

的改善,这样就可以通过减少训练数据来改善决策树的性能。John 提出了 Robust-C4.5 算法,该算法通过反复地训练决策树分类器和删除被当前决策树误分类的实例来减少训练数据,并提高决策树 C4.5 的性能。Brodley 等人提出的 CF 算法通过一致过滤器识别并删除误分类的训练数据,即先构建多个分类器,并使用它们识别误分类的训练数据,当且仅当所有的分类器都误分类了某个训练数据时,该训练数据才从训练集中删除。Sebban 等人提出了使用原型选择算法来减少原始数据而提高后剪枝决策树的性能。

聚类分析提供由个别数据对象到数据对象指派到簇的抽象,这些簇原型可以用作大量数据分析和数据处理技术的基础。唐南奇[7-8]等人验证了聚类抽取训练样本对 BP 神经网络在农用地分等中的有效性,提取的学习样本具有典型性。在相关领域的监督分类中聚类方法也能有效地抽取学习样本。这些文献表明,用聚类分析所选择的聚类中心点作为一个神经元的备选子集,当取较小的聚类标准和取较小删除标准时,所聚成的类的数目很多,这些类的中心便能够均匀地覆盖样本空间,使输入的样本均匀地覆盖在备选的样本空间中。

### 1.2 C5.0 算法简介

C5.0 决策树算法由 C4.5 算法改进而成,根据提供最大信息增益(Information Gain)的字段分割样本数据,并对决策树各叶子进行裁剪或合并来提高分类精度,最后确定各叶子的最佳阈值。通常不需花费大量的训练时间即可建立决策树,且生成的决策树容易进行解释。

下面以计算评价属性 $A$ 为例计算信息增益率 GainRatio $(A)$,$S$ 表示一组样本,$p_i$ 是任意样本属于 $B_i$ 的概率,用 $S_i/S$ 表示。假定类别属性具有 $n$ 个不同的值,定义 $n$ 个不同类 $B_i$ $(i=1,\cdots,n)$。设 $S_i$ 是类 $B$ 中的样本数。Info$(S)$表示当前样本中的信息熵:

$$Info(S)=\sum_{i=1}^{n} p_i \log(p_i) \tag{1}$$

设属性 $A$ 具有 $n$ 个不同值$\{A_1,A_2,\cdots,A_n\}$,利用 $A$ 将 $S$ 划分为 $n$ 个子集 $\{S_1,S_2,\cdots,S_n\}$,其中 $S_j$ 为 $S$ 在 $A$ 中具有 $A_j$ 的样本,$S_{ij}$ 是子集 $S_j$ 中类 $B_i$ 样本数。$Info(S,A)$ 表示利用属性 $A$ 划分 $S$ 中所需要信息:

$$Info(S,A)=\sum_{j=1}^{n} \frac{S_{1j}+S_{2j}+\cdots+S_{nj}}{S} Info(A) \tag{2}$$

分裂信息 $SplitInfo(A)$是 $S$ 关于属性 $A$ 的各值的熵,用以消除具有大量属性值属性的偏差,计算如下:

$$SplitInfo(A)=-\sum_{i=1}^{n} \frac{|S_j|}{|S|} \log\left(\frac{|S_j|}{|S|}\right) \tag{3}$$

$$Gain(A)=Info(S)-Info(A) \tag{4}$$

$$GainRatio(A)=Gain(A)/SplitInfo(S,A) \tag{5}$$

### 1.3 K-means 聚类算法简介

K-means 聚类算法又称为 $K$ 均值聚类算法,其优点是原理简单、算法速度快、伸缩性好。

K-means 聚类算法的工作流程是:首先随机选取 $K$ 个样本作为初始聚类中心,然后计算各个样本到聚类中心的距离,把样本归到离它最近的聚类中心所在的类中,重新计算调整后的新类的聚类中心,重复这个过程,直到相邻 2 次的聚类中心没有任何变化,这时样本调整结束,算法已经收敛。该算法的描述为:

(1)给定大小为 $N$ 的数据集,令 $I=1$,选取 $k$ 个初始聚类中心 $Z_j(1)$,$j=1,2,3,\cdots k$;

(2)计算每个数据对象与聚类中心的距离 $D(X_i,Z_j,(I))$。其中 $i=1,2,3,\cdots,n$,$j=1,2,3,\cdots,k$ 如果满足(6)式:

$$D(X_i,Z_k(I))=\min\{(X_i,Z_k(I)),i=1,2,3,\cdots,n\} \tag{6}$$

则 $X_i\in W_k$;

(3)计算 $K$ 个新的聚类中心可表示为:

$$Z_i(I+1)=\frac{1}{n}\sum_{i=1}^{nj} X_i^{(j)} \quad j=1,2,3,\cdots,k \tag{7}$$

(4)判断:若 $Z_i(I+1)\neq Z_i(I)$,$i=1,2,3,\cdots,K$,则 $I=I+1$,返回(2);否则该算法结束。

从上面的算法思想和算法框架,不难看出,$K$ 个初始聚类中心点的选取对聚类结果具有较大的影响,因为在该算法中是随机选取任意 $K$ 个点作为初始聚类中心。如果有先验知识,可以选取具有代表性的点作为初始中心点。

### 1.4 研究方法

考虑将 $k$ 个代表性的数据用于 C5.0 的模型训练,首先将数据集 $T$ 划分为 $K$ 个不相交的“类”,然后再从 $K$ 个类中的数据中心点附近随机抽取一个样本,这样就可以最终获得聚类采样数据子集,该学习样本更具典型性和代表性,实际效果较好。文中将最为广泛的 K-means 聚类分析抽取训练样本,通过减少后的训练数据提高 C5.0 决策树性能。

基于以上分析,研究方法具体流程如图 1 所示。

## 2 实例

有线电视服务交互服务是指有线电视台可利用现有的宽带网络和卫星传输系统,将海量的数据信息加密后广播出去,用户可按照自己的需要,通过无线遥控器、机顶盒、IC 卡等设备,在电视上自由地点播远程节目库中的视频节目和信息,以及股票信息接收、交互式娱乐、电子商务、电视购物、远程教育等增值业务服务。随着人们消费观念的转变,“有偿服务”也将成为未来信息服务产业的核心,付费电视业务将成为电视发展的动力和结构变化的方向,基于此进行的交互服务订制与否的用户预测具有广阔的市场前景。

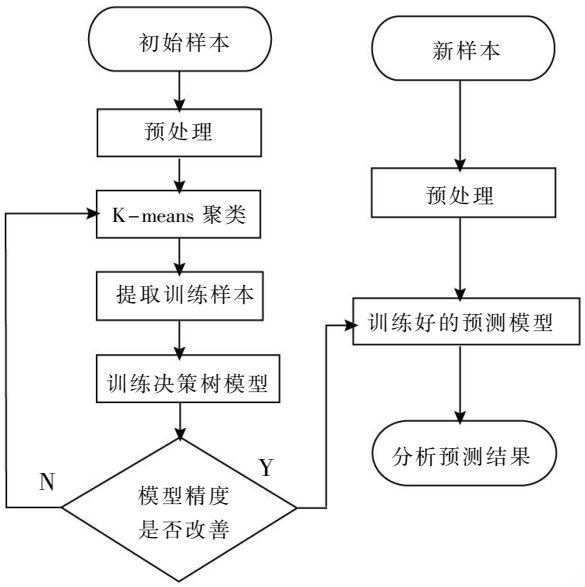


图 1 基于 K-means 聚类与决策树的有线电视交互服务订制预测建模框架与应用方法

### 2.1 数据准备与预处理

本文数据来源于某区有线电视网络的 442 名用户在一定历史时期市场研究资料,文中将 K-means 聚类分析与 C5.0 算法应用于有线电视交互服务的订制预测分析中,从中找出订制响应率较高的客户群,分析工具采用 SPSS Clementine 11.1。影响目标变量交互服务订制与否的因素很多,根据实际操作经验来看,通常包括:教育程度,性别,年龄,每天看电视时长,所属行业,子女数目,月收入等级等。预处理过程中包含对收入缺失值的补充,对月收入等级、所属行业进行离散化处理等。

### 2.2 抽取学习样本与建立决策树模型

决策树算法是从样本中学习规则,属于监督分类方法,因此学习样本的好坏对决策树模型的性能影响较大。本文依据渐进抽样原则,采用聚类分析中 K-means 算法对原始数据进行聚类抽样。为了保证评价模型的学习精度,学习样本的确定采用试验的方法,即根据建立的模型评价精度高低来选择合适规模的学习样本。

试验开始聚 50 个类,分别从每个聚类中心附近抽取一条记录,得到 50 个训练样本。通过建立评价模型来检验评价模型的精度,若满足实际情况的要求,表示该模型建立合理;如果所建立的评价模型精度不高,需要重新增加学习样本对模型进行训练,直至满足实际工作需要为止。依次增加学习样本数量,发现 150 个样本比 120 个样本评价模型准确率提高了 4.16 个百分点,而 200 个和 180 个学习样本模型的准确率只比 150 个样本模型提高了 0.56 个百分点和 0.38 个百分点。从这些数据可以看出在 150 个样本基础上每增加 30 左右的样本,模型精度提高的很少,表明在 150 个样本点模型准确率趋于稳定。因此确定 150 个学习样本建立的决策树模型作为有线电视交互服务订制预测模型,该样本所建立的模型预测精度为 87.35%。不同样本数量所得模型的精度如图 2 所示。

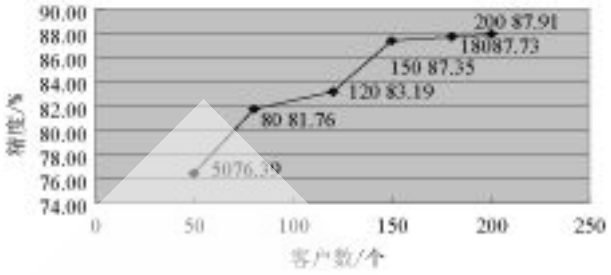


图 2 不同样本建立模型的精度

### 2.3 试验结果与分析

(1) 基于 K-means 与决策树 C5.0 预测的误分类损失。用 442 个样本替代流程图中的数据源,检验所建立的决策树模型,其模型测试的准确率为 84.84%,表明该模型的泛化能力较好,能有效地使用典型性样本推理出大量未知样本的类别。误分类损失如图 3 所示,字段 NEWSCHAN 表示实际的订制意愿,字段$C-NEWSCHAN 则表示预测的订制意愿,0 表示不愿意接受订制,1 表示愿意接受订制。

$C-NEWSCHAN

| NEWSCHAN | 0 | 1 |
|---|---|---|
| 0 | 199 | 28 |
| 1 | - 39 | 176 |

图 3 基于 K-means 与决策树预测的误分类损失

该图 3 表明,442 名用户中,共预测有 204 名用户会订制有线电视交互服务,其中准确预测 176 名,即愿意接受订制的预测准确率达 86.27%。预测有 234 名用户将不愿意订制有线电视交互服务,其中准确预测 199 名,即不愿意订制有线电视交互服务的预测准确率为 83.61%。

(2) 本文研究方法与基于决策树预测的误分类损失比较。同样的样本不经过聚类抽取学习样本而直接用决策树 C5.0 进行预测,误分类损失如图 4 所示。该模型的预测准确率仅达 70.14%,其中愿意接受订制的预测准确率为 77.48%,该精度与基于 K-means 与决策树进行预测所得的同类精度 86.27%低了 8.79 个百分点。

(3)基于 K-means 与决策树 C5.0 预测所产生的部分决策规则描述。在生成决策树之后,可以方便地提取决策树描述的知识,沿着根节点到叶节点的每一条对应一条决策规则。

抽取部分响应率较高的节点列如图 5 所示。

指数值大于 1 的节点表示,通过从这些节点中选

择记录而不是从整个样本中随机选择记录,能够有更多的机会找到愿意接受订制的用户。抽取上图第三条规则描述如下:

该用户群为 40 岁以上的女性,每天看电视 3 h 以内,这类用户在 442 名样本用户中共存在 63 位,占总体样本的 14.25%(14.25%=63×100%/442),63 位用户中有

| | $C-NEWSCHAN | |
|---|---|---|
| NEWSCHAN | 0 | 1 |
| 0 | 193 | 34 |
| 1 | 98 | 117 |

图 4 基于决策树预测的误分类损失

44 位用户愿意接受交互服务的订制,即响应率为 69.84%(44×100%/63),这 44 位用户数目占总体愿意接受订制数目(如图 3,39+176=215 名)的 20.47%(44×100%/215),从这些记录中获得积级响应的可能性是随机选择用户的 1.43 倍(总体响应率=215×100%/442=48.64%,1.43=69.84%/48.64%)。故可对响应指数>1 的用户群进行有针对性的营销。

| 规则 | 节点/个 | 节点/% | 收益/户 | 收益/% | 响应/% | 指数 |
|---|---|---|---|---|---|---|
| AGE<=40andGENDER<=0andEDUCATE>14 | 38 | 8.6 | 25 | 11.63 | 65.79 | 1.35 |
| AGE>40 and GENDER>0 and ORGS <=3 and INC <=2 and EDUCATE>11 | 27 | 6.1 | 23 | 10.7 | 85.19 | 1.75 |
| AGE>40 and GENDER<=0 and TVDAY <=3 | 63 | 14.25 | 44 | 20.47 | 69.84 | 1.43 |

图 5 生成分类树的部分节点收益表

实验结果表明,决策树建模前的样本抽样采用聚类抽样方法,以渐进抽样的原则,有效得减少了学习样本的数量,降低了评价模型的复杂度,并提高了模型的精度。

本文运用聚类方法抽取决策树模型的学习样本,有效地减少了学习样本空间,在试验精度不高时增加选择样本,所获得的模型的预测精度相对于仅运用决策树进行处理有所增高,并且该模型的可解释性较好,提取的规则能有效地识别高响应率客户群。进一步的工作包括对该方法进行完善,以及进行深入的理论上的分析和严格论证,在回归任务、神经网络、粗集等其他一些学习方法和其他集成学习方法的基础上进行评测。

**参考文献**

[1] SEBBAN M, NOCK R, CHAUCHAT J H, et al. Impact of learning set quality and size on decision tree performances [J]. IJCSS, 2000, 1(1): 85-105.

[2] 饶秀琪,张国基.基于 KPCA 的决策树方法及其应用[J].计算机工程与设计, 2007, 28(7): 1612-1613.

[3] DURKIN J,蔡竞峰,蔡自兴.决策树技术及其当前研究方向[J].控制工程, 2005, 12(1): 15-21.

[4] BRODLEY C E, FRIEDLM A. Identifying and eliminating mislabeled training instances [C]. Proc of the 13th National Conference on Artificial Intelligence, 1996: 799-805.

[5] OATEST, JENSEN D. The effects of training set size on decision tree complexity [C]. Proc of the 14th International Conference on Ma-chine Learning. Nashville, Tennessee: [s. n.], 1997: 379-390.

[6] YAROVOY A G, MATUZAS J, LEVITAS B,et al. UWB radar for human being detection [C]. Proc of International Radar Conference, 2005: 1-3.

[7] ZHANG Bu Han, ZHAO Jian Jian, LIU Xiao Hua. Short-termload forecasting based on wavelet neural network [J]. Power System Technology, 2004, 28(7): 15-18.

[8] 周志华,陈世福.神经网络集成[J].计算机学报, 2002,25(1):1-8.

(收稿日期:2009-06-17)

## 英特尔未来芯片:重构计算机,改写人机交互

2009 年 12 月 2 日,美国加州圣克拉拉--英特尔研究院的研究人员今天展示了一款处理器研究原型,在单芯片上实现了云计算机的功能,为笔记本电脑、PC 和服务器的设计方式提供了众多创新的设计理念。该项研究的长期目标是为未来计算机创建目前难以置信的扩展性能,促进开发全新的应用程序和人机界面。英特尔计划明年向行业和学术界合作伙伴提供 100 个以上的芯片原型用于实际研究,开发全新的软件程序和编程模式。

英特尔将在 2010 年初发布集成关键功能的新一代酷睿系列芯片产品,同年还将发布该系列的 6 核和 8 核处理器产品。而该芯片原型则拥有 48 个可完全编程的英特尔处理器内核,这也是有史以来集成度最高的单硅 CPU 芯片。它还整合了进行信息共享的高速片上网络,以及最新发明的电源管理技术,全部 48 个内核都实现了极高能效运行,其功耗可低至 25 瓦,运行最高能耗也仅为 125 瓦(与现在的英特尔处理器的能耗水平接近,仅相当于两个普通家用灯泡的耗电量)。

(英特尔公司供稿)