文章编号:1004-1478(2012)03-0049-04

基于信息熵的决策树挖掘算法在智能电力营销中的应用

戴小廷[1], 陈荣思[2], 肖冰[1]

(1.福建工程学院 交通运输系,福建 福州 350108;
2.福建省泉州电业局,福建 泉州 362000)

摘要:针对目前电力营销管理系统缺少有效的营销数据决策支持的问题,将基于信息熵的决策树挖掘算法应用于电力营销中,并建立电力客户分类模型.实际应用结果表明,该分类模型具有较好的预测分类能力,能够满足电力营销工作中的客户及时准确分类的需要.

关键词:信息熵;决策树;数据挖掘;电力营销
中图分类号:TM73 文献标志码:A

Application of decision tree mining algorithms based on information entropy in the intelligent electric power marketing

DAI Xiao-ting[1], CHEN Rong-si[2], XIAO Bing[1]

(1. *Dept. of Trans., Fujian Univ. of Tech., Fuzhou* 350108, *China*;
2. *Quanzhou Electr. Power Bearu of Fujian Province, Quanzhou* 362000, *China*)

Abstract: Aiming at the problem that the current electric power marketing management system was lack of effective marketing data for decision support, the decision tree mining algorithm was applied in electric power marketing based on the information entropy, and the power customer classification model was established. Practical application results indicated that the classfication model had better predictive classification ability, and was able to meet the needs of customers' timely and accurate classification in electric power marketing.

Key words: information entropy; decision tree; data mining; electric power marketing

0 引言

近年来电力市场深入发展,智能电网成为各国电力行业研究热点,有了智能电网的高级计量、IP通信、微网和自动闭环控制等技术的支持,用户可以更多地参与到电网运营中来[1-2].在这种环境下,电力公司的营销模式也应向智能营销转变,即从传统的仅考虑将电力安全地供给用户的模式到现在需要更多地了解电力市场中的客户对象、分析其消费行为并与其实时互动的营销模式转变.而其中开

收稿日期:2011-12-15
基金项目:福建工程学院2011年度科研发展预研基金(GY—Z11077)
作者简介:戴小廷(1972—),男,江西省宜春市人,福建工程学院副教授,主要研究方向为信息管理与信息系统、物流管理.

展营销数据的智能分析是智能电力营销中关键的一环.

在电力营销的实际工作中,建立了诸如电力营销管理信息系统、客户服务支持系统等众多的信息系统,积累了海量的企业营销运营数据. 但是信息化建设时间、平台和开发商的差异,造成了数据异构性强、数据冗余、多数据源等问题,使得系统难以提供高效智能的营销决策支持. 一方面是数据资源的大量闲置,另一方面智能电力营销管理又迫切要求更加有效的营销数据决策支持,提高运行效率. 近几年,电力系统营销方面的数据挖掘研究取得了不少令人欣喜的成果[3-10],但是数据挖掘技术在智能电力营销中的作用还没有完全发挥,还缺乏比较成熟的能够在整体上推广的涵盖各种数据挖掘技术的软件,离全面的营销决策支持需求还是有一定距离. 本文主要探讨基于信息熵(entropy)的决策树挖掘算法在电力营销中的具体应用.

1 基于信息熵的决策树挖掘算法

1.1 基于信息熵的决策树挖掘算法原理

信息熵又称为期望信息量,是用来衡量信息量凌乱程度(不确定性)的指标,熵值愈大,则代表信息的凌乱程度愈高. 基于信息熵的决策树挖掘算法是通过收集已知类别的样本,将提供最大信息增益的属性作为节点分裂方案去构造决策树,即所选的测试属性是从根到当前节点的路径上尚未被考虑的具有最高信息增益的属性. 决策树的每个节点对应一个非类别属性,每条边对应该属性的每个可能值[11].

评估属性 A 相对于范例集合 U 的信息增益(互信息)被定义为

$$I(A) = H(U) - H(U|V) \quad ①$$

其中

$$H(U) = -\sum_{i=1}^{s} P(U_i)\log_2 P(U_i) \quad ②$$

$$H(U|V) = -\sum_{k=1}^{m} P(V_k)\sum_{i=1}^{s} P(U_i/V_k)\log_2 P(U_i/V_k) \quad ③$$

1.2 基于信息熵的决策树挖掘过程

1) 选择训练样本,将其分成属于类别 $U_i(i=1,2,\cdots,s)$ 的样本集合,计算出 $P(U_i)$.

2) 将 $P(U_i)$ 代入式②,计算初始熵 $H(U)$.

3) 对每个属性 $A_k(k=1,2,\cdots,m)$ 按其属性值,计算其相应的概率 $P(V_k)$,$P(U_i/V_k)$.

4) 由③得到实体类别的平均不确定性 $H(U|V)$.

5) 依据式①计算由属性 A_k 引起的信息增益 $I(A)$.

6) 从 A_k 中选择最大信息增益的属性作为根节点,该属性的每个值作为该节点的一个分支,重复1—5步,利用训练样本生成决策树模型,直到所有的样本都归为一个分支并且所有叶节点只包含一类时,终止计算,得到决策树.

7) 对决策树进行修剪,去除噪音或者异常的数据.

8) 利用生成的决策树对未知数据进行分类,获取有益的决策支持信息.

1.3 建树阶段(伪代码)[11]

```
Partition (T);
Partition (Data S)
If (all points in S are in the same class) then return;
Evaluate splits for each attribute A
Use best split found to partition S into S1 and S2;
Partition (S1);
Partition (S2);
```

2 应用实例

2.1 基于信息熵的电力客户分类决策树构建

地市级供电企业是省电力公司的分公司,因此其主要经营目标是在保障安全供电的基础上完成省公司下达的电量增长、平均电价、电费回收、销售收入等经济指标. 因此,对客户的分类管理应从有利于经济指标完成入手. 在对某地市级客户资料分析的基础上,根据营销实际,将客户的类别分为重要客户(A 类客户)、潜力客户(B 类客户)、普通客户(C 类客户)3 类. 根据客户的社会重要性、用电量大小、供电安全可靠性要求、电价高低、客户信用等主要特性,从客户数据库中选取部分数据经过清洗、转换,去除了噪声数据和不一致性之后,进行概化处理,如将社会重要性分为重要、中等、一般、低 4 个等级,用电量大小概化为大、中、小 3 个等级,供电安全可靠性要求、电价高低概化为高、中、低 3 个等级,将客户信用概化为高、中、低 3 个等级. 抽取其中部分数据作为训练集,得到开展数据挖掘的初始数据,见表 1.

表 1　数据挖掘初始数据

序号	社会重要性	用电可靠性	用电量	电价高低	客户信用	客户分类
1	重要	高	大	低	高	A
2	中等	高	中	低	中	A
3	一般	中	大	低	高	A
4	一般	高	小	高	低	B
5	一般	中	中	低	中	C
6	一般	高	中	低	中	B
7	重要	中	大	中	高	A
8	一般	高	大	高	高	A
9	一般	中	大	高	高	A
10	一般	中	大	低	高	A
11	中等	高	大	低	高	A
12	中等	中	中	低	高	B
13	一般	高	大	低	低	B
14	一般	高	小	低	高	C
15	中等	中	中	低	低	C
16	重要	高	大	中	低	A
17	一般	低	中	高	高	C
18	一般	低	中	低	高	C
19	一般	中	小	中	高	B
20	低	高	大	低	中	C

基于信息熵的决策树挖掘过程如下:

1) 客户分类 U 中有 20 个例子数,其中 9 个 $U_1(A)$,6 个 $U_2(B)$,5 个 $U_3(C)$.

$$H(U) = -\sum_{i=1}^{s} P(U_i)\log_2 P(U_i) =$$

$$-\frac{9}{20}\log_2 P\left(\frac{9}{20}\right) - \frac{6}{20}\log_2 P\left(\frac{6}{20}\right) -$$

$$\frac{5}{20}\log_2 P\left(\frac{5}{20}\right) = 1.539\ 5$$

2) 属性"社会重要性"的取值分为 V_1 = 重要,V_2 = 中等,V_3 = 一般,V_4 = 低.

$$H(U|V) = -\frac{3}{20}\left[\frac{3}{3}\log_2(3/3)\right] -$$

$$\frac{4}{20}\left[\frac{2}{4}\log_2(2/4) + \frac{1}{4}\log_2(1/4) + \frac{1}{4}\log_2(1/4)\right] -$$

$$\frac{12}{20}\left[\frac{4}{12}\log_2(4/12) + \frac{5}{12}\log_2(5/12) + \frac{3}{12}\log_2(3/12)\right] -$$

$$\frac{1}{20}\left[\frac{1}{1}\log_2(1/1)\right] = 1.232\ 8$$

3) 计算"社会重要性"属性的信息增益.

$$I(\text{社会重要性}) = H(U) - H(U|V) =$$

$$1.539\ 5 - 1.232\ 8 = 0.306\ 7$$

同理得到用电量、用电可靠性、电价高低、客户信用等属性的信息增益分别为: I(用电可靠性) = 0.267 2; I(用电量) = 0.340 8; I(电价高低) = 0.135 4; I(客户信用) = 0.108 8.

4) 比较上述信息增益,选取最大信息增益的属性"用电可靠性"作为决策树根节点,属性值作为叶节点,对每一个叶节点,考虑其他属性,循环上述计算步骤,得到客户分类决策树,如图 1 所示.

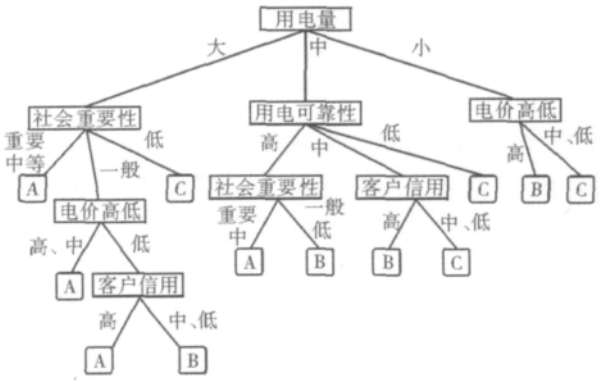

图 1　客户分类决策树

2.2　决策树的知识表示

根据数据挖掘得到的客户分类决策树,采用 IF – THEN的格式可以得到各种分类规则,如:

1) IF 用电量 = "大"AND 社会重要性 = "重要" OR "中等",THEN 客户分类为 A 类.

2) IF 用电量 = "大"AND 社会重要性 = "一般" AND 电价高低 = "高"OR "中等",THEN 客户分类为 A 类.

3) IF 用电量 = "大"AND 社会重要性 = "一般" AND 电价高低 = "低"AND 客户信用 = "中"OR "低",THEN 客户分类为 B 类.

4) IF 用电量 = "大"AND 社会重要性 = "低", THEN 客户分类为 C 类.

将决策树中得到的类似以上规则的所有分类规则储存在企业智能营销知识库中,即可应用在实际电力客户营销工作中.

2.3　分类模型的评估与检验

选取供电企业中部分实际业务数据(100 个样本),对上述数据挖掘基础上得到的分类规则进行检验. 分类准确率为正确的分类样本数量占检验样本总数量的百分数. 通过验证,模型的分类准确率为 97%,具有较好的预测分类能力,能满足智能电力营销工作中的客户及时准确分类实际工作的需要.

2.4 知识在智能营销工作中的具体应用

在上述数据挖掘基础上得到的分类规则,电力营销工作人员在客户关系管理中就可以通过知识库系统来使用它,比如可以根据规则判断一个新客户应该属于哪一种类型,从而采取有针对性的营销策略,如针对 A 类客户公司可以指派专门的营销人员负责联络,定期走访,给他们提供用电价格、用电知识、可靠性等方面快捷周到的服务,平时营销人员更多关注该客户所在行业动态,了解国家对该行业的政策,优先处理该类客户的抱怨和投诉等. 通过细分市场需求,将重要客户、潜力客户、一般客户进行区分,便于优化配置企业营销资源,给用户更多有针对性的服务,促进电力市场的发展.

3 结语

数据挖掘等信息手段应用于智能电力营销中,可以使电力公司及时、准确、快速地洞悉客户相关信息,开展客户关系管理,实施差异化服务策略,进一步开拓市场,满足智能化电力市场的需求. 本文将基于信息熵的决策树挖掘算法应用于智能电力营销中,并建立电力客户分类模型. 实际应用结果表明,该分类模型具有较好的预测分类能力,能满足智能电力营销工作中的客户及时准确分类实际工作的需要. 但这仅是智能电力营销中的小部分内容,今后需要拓展这类研究,为智能电力营销实际工作提供更多更好的决策支持.

参考文献:

[1] 胥威汀,刘俊勇,刘友波,等. 智能营销研究概述(一)——国外研究现状与启示[J]. 电力自动化设备,2010,30(2):139.

[2] 姚珺玉,刘俊勇,刘友波,等. 智能营销研究概述(二)——我国智能营销发展战略与机遇[J]. 电力自动化设备,2010,30(3):129.

[3] 刘文霞,王志强,毛辉,等. 数据集成技术在电力营销数据分析系统中的应用[J]. 电网技术,2004,28(18):70.

[4] 方兆本,杨培洁,彭甘霖,等. 电力客户信用风险实证研究[J]. 电力系统自动化,2005,29(1):61.

[5] 张珂,王玉凡,苑津莎,等. 基于云模型和关联分析法的电力营销目标市场模糊评价[J]. 华北电力大学学报,2009,36(4):30.

[6] 王志刚,曲巍,黄爱颖. 天津市电力公司售电市场实时预测分析系统建设及应用[J]. 电力需求侧管理,2007,9(1):48.

[7] 侯勇,张荣乾,谭忠富,等. 基于模糊聚类和灰色理论的各行业与全社会用电量关联分析[J]. 电网技术,2006,30(2):46.

[8] 王羽,杨道辉,马光文,等. 因素分析法在电力营销分析中的应用[J]. 水力发电,2010,36(4):4.

[9] 肖峻,王成山,周敏. 基于区间层次分析法的城市电网规划综合评判决策[J]. 中国电机工程学报,2004,24(4):50.

[10] 杨滋荣,陈建中,文静华. 分布式数据挖掘在电力客户系统中的应用[J]. 电力系统及其自动化学报,2007,19(4):23.

[11] 郭迎春. 知识型电力客户关系管理研究[D]. 北京:华北电力大学,2008.