基于决策树的数据挖掘算法应用研究

常秉琨，李 莉
（郑州职业技术学院，河南 郑州 450121）

摘 要：以决策树数据挖掘分类算法在金融客户关系管理（CRM）中的应用为例，进行了数据挖掘的尝试，从中发现企业产品的销售规律和客户群特征，从而提高CRM对市场活动和销售活动的分析能力，得到了与实际经验相符的结果和相应的“规则”，验证了其可行性和可供决策支持的现实意义。

关键词：算法；数据挖掘；决策树；客户关系管理
中图分类号：TP311 **文献标识码**：A

Applied research of data mining algorithm based on decision tree

CHANG Bing Kun, LI Li
(Zhengzhou Technical College, Zhengzhou 450121,China)

Abstract: This article takes the decision tree data mining sorting algorithm's application in the financial customer relations management (CRM) as an example, to carry on the data mining attempt, and discover the regulations of enterprise product sales and customer base characteristics, thereby enhance the CRM on the analytical capacity of market activities and sales activities, and obtain the result which consistent with the practical experience and corresponding“the rule” to verify its feasibility and the decision support for practical significance.

Key words: algorithm; data mining; decision tree; customer relationship management

决策树技术是一种对海量数据集进行分类的非常有效方法，通过决策树的构造模型,从海量信息中挖掘有效的数据,提取有价值的分类规则,从而获得有用的知识,为决策者提供支持，帮助决策者准确地预测[1]。本文研究了基于决策树的数据挖掘的相关理论发展及实际应用，尤其是在商业中的应用，研究了决策树算法在数据挖掘中应用，给出了在金融客户关系管理中的具体算例。

1 决策树的分类挖掘技术

1.1 决策树的分类挖掘算法

分类是数据挖掘中应用最多的任务，要为每个类别做出准确的描述或建立分析模型或挖掘出分类规则，然后用这个分类规则对其他数据库中的记录进行分类[2]。在具体分类中，总是希望进行较少的属性测试，较快地给实例分类，因此在构建决策树时，树的高度越小越好。对于N个样本，它们分成属于类别C_i（$i=1,2,\cdots,C$)的样本集合,类别C_i中的样本个数为N_i，每个样本有K个属性，每个属性有J_K个值。决策树的构造过程如下：

(1)计算初始熵(熵用字符S表示):

$$S(I)=\sum_{i=1}^{c}-(N_i/N)\mathrm{Log}_2(N_i/N)=\sum_{i=1}^{c}-P_i\mathrm{Log}_2P_i$$

（2）选择一个属性作为决策树的根节点：

①对每个属性A_k（$k=1,2,\cdots,K$）按照属性A_k的J个A_{kj}值，把原始样本分成第1级样本集。虽然A_{kj}的分支含有n_{kj}个样本，但它们不一定属于单一的类别；

②对于每个分支的n_{kj}个样本，属于类别C_i的样本数目是$n_{kj}(i)$，用下式可以求出该分支的熵：

$$S(I,A_k,J)=\sum_{i=1}^{c}-(n_{kj}(i)/n_{kj})\mathrm{Log}_2(n_{kj}(i)/n_{kj})$$

$$S(I,A_k)=\sum_{j}\sum_{i}^{c}-(n_{kj}/N)\times\left[-n_{kj}\mathrm{Log}_2(n_{kj}(i)/n_{kj})\right]$$

③计算由测试属性引起的熵降低，即 $\Delta S(K)=S(I)-S(I,A_K)$；

④选择产生最大熵降低的属性 A_{K0}，即 A_{K0} 满足：$\Delta S(K_0)>\Delta S(K)$($K$= 1 ,2 , …, K ,且 $K\neq K_0$)

⑤属性 A_{K0} 便是决策树的根。

(3)由属性将产生 J_{K0} 个叶节点，并将样本集分成 J_{K0} 个子集，对每个叶节点上的样本子集依次利用上面的方法选择一个属性 Ay 作为决策树的下一级，使在该叶节点能得到最大的熵降低。

(4)按照步骤(3)不断构造决策树的下一级直至所有的样本子集只有一个类别，这时表明系统的熵为零，决策树构造过程完毕。

1．2 决策树分类挖掘系统的建立

在对金融客户进行分类分析的过程中，决策树分类数据挖掘系统建立和应用的一个典型过程是：(1)根据客户分类的标准，执行客户分类算法，并将运行结果存储于数据仓库中，这样，每个现有的客户都具有一个确定的客户类别；(2)根据历史数据，主要是客户的背景数据和客户的分类数据，执行决策树生成算法，针对每一种客户类别，生成一棵决策树，以一定的形式存放于数据仓库中；(3)在以上过程执行完毕后，当一个新的客户来办理业务时，客户经理可以首先在系统中调用决策树展示模块，系统将整个决策树展示出来，然后系统根据客户的具体背景情况预测客户所属的客户类别，以及属于该客户类别的概率，并将这些情况展现给客户经理，客户经理根据这些情况，对该客户采取相应的营销策略，从而达到较好的效果。

2 基于决策树的数据挖掘的案例分析

2.1 公司客户关系数据库

利用决策树实现客户细分，主要是在基于客户价值的客户细分方法之上进行的，目的是通过了解客户的特征性指标和行为性指标与客户所在客户类别的关系，可以了解同一价值客户的差异性，有针对性地对不同客户制定相应的销售策略[3]。下面以河南省某金融企业的客户关系管理（CRM）为例，该公司 CRM 数据如表 1 所示，说明基于决策树的客户分类数据挖掘技术在金融企业客户关系管理中的应用，来具体阐述如何运用改进 ID3 算法进行数据挖掘。由于是以分析客户的特征性指标为例的，所以从中选取了代表特征性的 3 个属性：客户年龄段、学历以及职业。

表1 某公司CRM数据库

ID	姓名	职业	年龄	学历	客户类别
1	黎　明	企业员工	27	高等教育	Ⅳ类客户
2	张　红	工　人	22	初等教育	Ⅱ类客户
3	王　榆	技术工程师	45	高等教育	Ⅳ类客户
4	李书亭	作　家	28	高等教育	Ⅳ类客户
5	侯建国	退休员工	60	中等教育	Ⅲ类客户
6	李小文	国税局科员	23	高等教育	Ⅱ类客户
7	段　军	记　者	30	高等教育	Ⅳ类客户
8	刘　星	销售经理	50	初等教育	Ⅰ类客户
⋮	⋮	⋮	⋮	⋮	⋮
100	李　林	公务员	43	高等教育	Ⅳ类客户

2.2 基于 ID3 的细分步骤

对于 N 个样本，它们分成属于类别 C_i（i = 1 ,2 , …, C)的样本集合,类别 C_i 中的样本个数为 N_i，每个样本有 K 个属性，每个属性有 J_k 个值。类别是客户类别，分为Ⅰ类客户、Ⅱ类客户、Ⅲ类客户和Ⅳ类客户 4 类。将具体的客户年龄概化为≤ 25 、25 ~ 50 和≥ 50 3 个年龄段，按学历分为初等学历、中等学历和高等学历 3 类。按职业分为商业人员、企业人员和其他人员 3 类。分别如表 2、表 3、表 4 所示。

表2 客户类别表

$C1$	$C2$	$C3$	$C4$
Ⅰ类客户	Ⅱ类客户	Ⅲ类客户	Ⅳ类客户

表3 属性表

属性（A_i）	K=1	K=2	K=3
客户属性	客户年龄段	学历	职业

表4 属性值表

属性值	$J1$	$J2$	$J3$
客户年龄段	≤25	25～50	≥50
学历	初等教育	中等教育	高等教育
职业	商业人员	企业人员	其他人员

决策树的构造过程如下：

(1)计算初始熵(熵用字符 S 表示)：

$$S(I)=\sum_{i=1}^{c}-(N_i/N)\mathrm{Log}_2(N_i/N)=\sum_{i=1}^{c}-P_i\mathrm{Log}_2P_i$$
$$=-1/100\log_2(1/100)=1/100\times10=1/10 \quad (1)$$

(2) 选择一个属性作为决策树的根节点：

①对每个属性 A_k（k = 1 ,2 , …, K）按照属性 A_k 的 J 个 A_{kj} 值，把原始样本分成第 1 级样本集。虽然 A_{kj} 的分支含有 n_{kj} 个样本，但它们不一定属于单一的类别。

②对于每个分支的 n_{kj} 个样本，属于类别 Ci 的样本数目是 n_{kj} (i)，用下式可以求出该分支的熵：

$$S(I,A_k,J)=\sum_{i=1}^{c}-(n_{kj}(i)/n_{kj})\ \mathrm{Log}_2(n_{kj}(i)/n_{kj}) \quad (2)$$

$$S(I,A_k)=\sum_{j=1}^{j}\sum_{i=1}^{c}-(n_{kj}/N)\times\left[-n_{kj}\mathrm{Log}_2(n_{kj}(i)/n_{kj})\right] \quad (3)$$

计算过程如下：

$S(I,A_1)=-1/32\log_2(1/32)=1/32\times5=5/32$

$S(I,A_2)=-1/52\log_2(1/52)=1/52\times5.6=14/130$

$S(I,A_3)=-1/16\log_2(1/16)=1/16\times4=1/4$

③计算由测试属性引起的熵降低，即

$$\Delta S(K)=S(I)-S(I,A_K)$$

$$S(I,A_1)-S(I)=5/32-1/10=9/160$$

$$S(I,A_2)-S(I)=14/130-1/10=1/130$$

$$S(I,A_3)-S(I)=1/4-1/10=3/20$$

④选择产生最大熵降低的属性 A_{K0}，即 A_{K0} 满足：

$$\Delta S(K_0)>\Delta S(K)(K=1,2,\cdots,K,且K\neq K_0) \quad (4)$$

3/20 > 9/160 > 1/130

根据以上计算结果，得出最大熵降低的属性是K3，即职业。

⑤属性职业便是决策树的根。

(3) 由属性将产生 J_{K0} 个叶节点，并将样本集分成个子集，对每个叶节点上的样本子集依次利用上面的方法选择一个属性 A_y 作为决策树的下一级，使在该叶节点能得到最大的熵降低。

(4) 按照步骤(3)不断构造决策树的下一级直至所有的样本子集只有一个类别，这时表明系统的熵为零，决策树构造过程完毕。然后根据上述的决策树构造过程，得到图1所示的决策树。

2.3 具体应用分析

决策树算法中属性的取值，进一步细分所依据的是特征性指标和行为性指标，然后可以分别得出各种特征性指标和行为性指标与4类客户之间的树状分类结构。下面以分析特征性指标为例，来说明决策树的构建。

决策树构建之前，必须要找出决策树的主属性。决策树主属性的确定主要是根据具体的情况而定。所以客户细分的主属性应该是“客户类别”，就是基于客户价值的客户细分得出的客户类别。

对于数据的选取，并不是所有的数据都符合要求，决策树建立所要求的数据应是没有噪音数据和缺失数据，这就需要对数据进行汇总处理。汇总处理一方面是将企业不同部门和不同分销机构的数据进行集成；另一方面是将数据进行概化处理，即将低层次的原始数据替换为高层次的概念，以便于进行数据挖掘。

分析图1，从中可以明确4类客户的特征属性的大致排布情况，根据所了解的情况，能够很方便地从客户的这些特征中大概了解其在4类客户中所处的位置。用IF－THEN的格式来表示树状图的信息(以第Ⅲ类客户为例)，例如：

IF 职业＝“商业员工”AND 年龄> 25 AND 年龄< 50 AND 学历＝“初等教育”THEN 属于Ⅲ类客户

IF 职业＝“企业员工”AND 年龄> 25 AND 年龄< 50 AND学历＝“初等教育”OR 学历＝“中等教育”THEN 属于Ⅲ类客户

IF 职业＝“其他员工”AND 年龄> 25 AND 学历≠“高等教育”THEN 属于Ⅲ类客户

这些式子说明：对于“商业员工”的客户来说，如果年龄在25～50岁之间，学历为“初等教育”，属于Ⅲ类客户；对于“企业员工”的客户来说，如果年龄在25～50岁之间，学历为“中等教育”或者“高等教育”，属于Ⅲ类客户；对于“其他员工”的客户来说，如果年龄大于25岁，只要学历不是“高等教育”，就属于Ⅲ类客户。

根据上面分析第Ⅲ类客户的结果，可以得出：不论客户职业是什么，只要年龄在25～50岁之间，学历为“初等教育”的客户，都属于Ⅲ类客户。由于Ⅲ类客户是企业目前利润的重要来源，所以企业就要吸引和保持住与Ⅲ类客户之间的客户关系，需要把营销策略更倾向于年龄在25～50岁之间的客户，而对于在该区间之外的客户，需要有选择性地开展促销方式，例如对其他职业的客户，就要考虑大于50岁的情况。由于Ⅰ类客户对企业的贡献很小，分析的必要性不大，所以重点是对Ⅱ、Ⅲ和Ⅳ类客户的分析。通过这种分析方式，逐步把通过ID3得出的树状结构的所有分支都进行分析，可以了解企业的同一价值客户在特征属性上的差异性。同理再对客户的行为性指标进行分析，可以得出同一价值客户在行为属性上的差异性。结合这两方面，就能

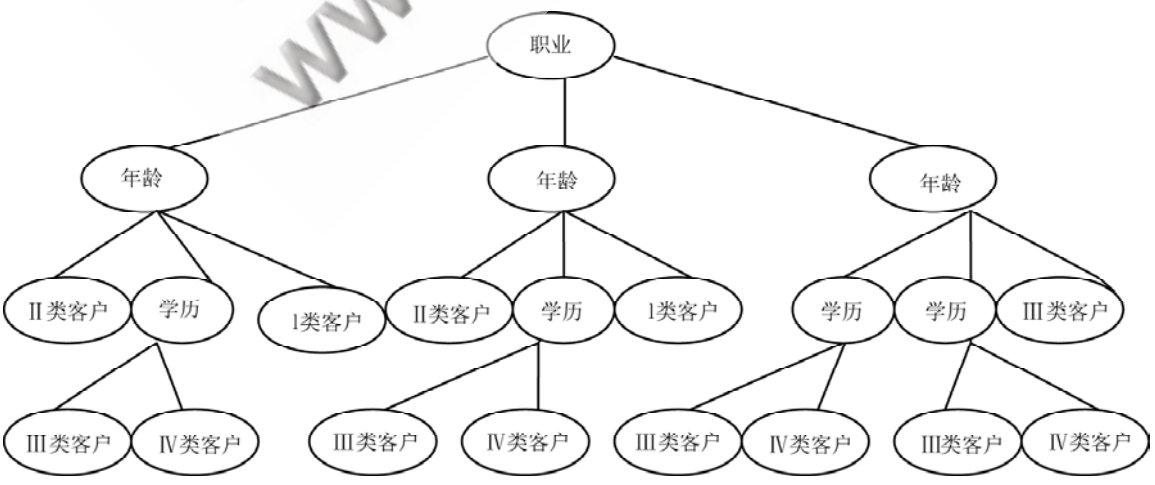

图1 决策树

(下转第68页)

(上接第64页)

够在了解客户价值类别的基础上，有针对性地对不同客户制定相应的销售策略，减少企业不必要的开销，实现对客户价值细分后的进一步细分。

如何高效地整合和分析企业各部门和各级分销机构内的销售和客户信息，使企业能够从全局的角度了解和认识市场是CRM的重要任务之一，而基于决策树的数据挖掘算法对于企业来说刚好可以实现这个任务。通过基于决策树的客户分类数据挖掘技术，可以了解客户的特征性指标和行为性指标与客户所在客户类别的关系[4]，使企业能够在了解何种资源组合可以使得自己获得高利润的同时，有针对性地根据客户的差异化和多变性需求，制定相应的销售策略，使得企业制定的营销策略更加符合市场的需求，保持在市场中的竞争地位。

参考文献

[1] 张世海，刘晓燕，涂庆，等. 基于决策树的高层结构智能选型知识发现[J]. 哈尔滨工业大学学报，2005，37(4):451-454.

[2] 崔立新，苑森森，赵春喜.约束性相联规则发现方法及算法[J].计算机学报,2000,22(2):216-220.

[3] 李绪成,王保保.挖掘关联规则中Apriori算法的一种改进[J].计算机工程与应用,2002,28(7):104-105.

[4] 魏定国,彭宏.基于知识网络的数据挖掘[J].计算机科学,2006,33(6):207-210.

(收稿日期：2009-02-22)