基于决策树的不完整数据的处理

张 靖，姚 珍，唐雪飞

(攀枝花学院网络中心 四川 攀枝花 617000)

【摘要】基于数据采集过程中常常存在一些不完整数据，以及不完整数据总是和样本空间中其他完整数据存在一定的相似性，提出了一种基于决策树的不完整数据的处理方法。在对不完整数据处理方法的比较、分析的基础上，采用一种有效的决策树方法对不完整数据处理。实例验证证明该决策树方法在不完整数据处理方面有很好的效果。

关 键 词 决策树; 不完整数据; 处理

中图分类号 TP311.13 文献标识码 A

Treatment of the Incomplete Data Based on the Decisive Tree

ZHANG Jing，YAO Zhen，TANG Xie-fei

(Campus Network Center, Panzhihu University Panzhihua Sichuan 617000)

Abstract The incomplete data, came into being in the process of data collection, always have some similarities to the complete data in the sample space. This article presents a treatment of the incomplete data based on the decisive tree. On the base of the comparison and analyses of the treatment of the incomplete data, a useful decisive tree is used to deal with such data, and at last, it is testified by an experiment. The experiment shows that the decisive tree can achieve a good result in the treatment of the incomplete data.

Key words decision tree; incomplete data; treatment

高质量的数据是数据挖掘成功的前提条件[1-2]。如果数据库和数据仓库中的数据量太大，其中必然会存在不完整的、含噪声的和不一致的数据。数据的不完整是影响数据质量的一个重要因素[1]。高质量的决策必然依赖于高质量的数据[3]，因此，必须对数据进行预处理。本文运用数据挖掘中决策树的思想，提出了一种决策树处理方法，对不完整数据进行处理，并对不完整数据的空缺属性值进行预测。

1 决策树方法

决策树方法是应用于数据挖掘分类问题的一种方法。该方法从一组无秩序、无规则的事例中推理出决策树表示形式的分类规则。决策树是一个类似于流程图的树结构，树结构中的每个内部节点代表一个属性上的测试；每个分枝代表一个测试输出；每个树叶节点代表一个类。所以从决策树的根到叶结点的一条路径就对应着一条取舍规则，整棵决策树就对应着一组析取表达式规则。因此能在一个样本的某些属性已知的情况下预测另外一个样本的未知属性。

决策树方法是广泛运用于数据挖掘分类问题的一种方法[4]，目前有多种决策树方法，如CART、CHAID、ID3、C4.5、SLIQ、SPRINT、C5.0等。本文运用的是决策树算法中的ID3算法。

ID3决策树学习是一种逼近离散值目标函数的方法，是运用得最广泛的一种归纳学习方法。

设属性A具有v个不同值$\{a_1,a_2,\cdots,a_s,\cdots,a_v\}$。可以用$A$将集合$S$划分为$v$个子集$\{S_1,S_2,\cdots,S_j\cdots S_v\}$，其中的$S_j$包含$S$中在$A$上具有值$a_j$的这样一些样本。设$s_{ij}$是子集$S_j$中类$C_i$的样本数，由$A$划分成子集的熵由下式表示：

$$E(A)=\sum_{j=1}^{v}\frac{S_{1j}+S_{2j}+\cdots+S_{mj}}{S}I(S_{1j},S_{2j},\cdots,S_{mj})$$

式中 $\frac{S_{1j}+S_{2j}+\cdots+S_{mj}}{S}$ 充当第j个子集的权，并且等于子集(即A的值为a_j)中的样本个数除以S中的样本总数。熵值越小，子集划分的纯度越高。对于给定的子集S_j有：

$$I(S_{1j},S_{2j},\cdots,S_{mj})=-\sum_{i=1}^{m}p_{ij}\log_2(p_{ij})$$

式中 $P_{ij}=S_{ij}/|S|$ 是S_j中的样本属于类C_i的概率。

收稿日期：2006 - 08 - 10

作者简介：张 靖(1972 –)，男，硕士，主要从事软件工程和计算机网络方面的研究.

属性A上的信息增益为：

$$G_{\text{Gain}}(A)=I(S_1,S_2,\cdots,S_m)-E(A)$$

ID3算法计算每个属性的信息增益，并选取具有最高增益的属性作为给定集合S的预测属性。对被选取的测试属性创建一个结点，并以该属性标记；对该属性的每个值创建一个分支，并据此划分样本。

2　决策树在不完整数据处理中的应用

对于数据源中不完整数据的清理，首先要检测数据源中的不完整数据，然后判断数据的可用性。对确认要保留的不完整数据记录采用一定的方法来处理记录中丢失的属性值[5]。一般常用的方法有：

(1) 人工填写空缺值，该方法很费时，特别是数据集很大、缺少值很多时，一般行不通；(2) 使用一个全局常量填充，如对所有属性值用同一个常量来填充，当空缺值的属性名和这个常量同名时将导致错误的分析结果；(3) 使用属性的平均值填充，如工资表中的工资空缺值可以用全部职工的平均工资来填充；(4) 使用与给定元组属于同一类的所有样本的平均值；(5) 使用最可能的值填充，如回归、线性预测等。

对不完整数据的处理方法很多，但是大多数处理方法都过于简单，没有考虑到整个数据空间的相似性。本文运用决策树的方法对不完整数据进行处理，其基本思想是尽量使整个样本空间的相似度增大。首先判断数据的可用性，然后根据已有的完整数据产生决策树，再把不完整数据中的已知属性输入决策树，最后预测不完整数据中未知属性的值。这种根据已知属性预测未知属性的方法得到的值，与整个数据空间其他数据的相似性很强。

算法的具体步骤如下：(1) 依次判断每个不完整数据的可用性，这一步骤可根据实际进行判断；(2) 统计不完整数据中的空缺属性数目k，记各属性为$C_i(i=1,2,\cdots,k)$；(3) 将完整数据输入决策树生成算法ID3，生成一棵以C_i为叶子节点的决策树；(4) 输入所有空缺属性为C_i的数据点，得到空缺属性C_i的值；(5) 重复步骤(3)，直到$i=k$。

3　实例

表1给出了一组不完整采样数据，该组不完整采样数据有5个属性，分别为C_1、C_2、C_3、C_4、C_5。由该表可以看出，其中C_1和C_5为不完整属性。下面将要以其中的完整数据推断出不完整属性的值。首先采用决策树推断预测属性C_1的值，决策树以C_1为叶子节点，分别计算C_2、C_3、C_4、C_5的信息增益$G_{\text{Gain}}(C_2)$ = 0.316 7、$G_{\text{Gain}}(C_3)$ = 0.048 54、$G_{\text{Gain}}(C_4)$ = 0.031 4、$G_{\text{Gain}}(C_5)$ = 0.483 3，其中C_5的信息增益最大，故以C_5作为根节点。再分别计算每个分支中余下属性的信息增益，得到如图1所示的决策树。

表1　不完整数据集

索引	C_1	C_2	C_3	C_4	C_5
1	$C_1\_v_1$	$C_2\_v_1$	$C_3\_v_1$	$C_4\_v_1$	$C_5\_v_1$
2	$C_1\_v_1$	$C_2\_v_1$	$C_3\_v_1$	$C_4\_v_2$	$C_5\_v_1$
3	$C_1\_v_1$	$C_2\_v_1$	$C_3\_v_1$	$C_4\_v_3$	$C_5\_v_1$
4	$C_1\_v_1$	$C_2\_v_2$	$C_3\_v_2$	$C_4\_v_2$	
5	$C_1\_v_1$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_1$	$C_5\_v_2$
6	$C_1\_v_2$	$C_2\_v_1$	$C_3\_v_1$	$C_4\_v_3$	$C_5\_v_2$
7	$C_1\_v_3$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_1$	$C_5\_v_1$
8		$C_2\_v_2$	$C_3\_v_2$	$C_4\_v_1$	$C_5\_v_1$
9	$C_1\_v_3$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_3$	$C_5\_v_1$
10	$C_1\_v_3$	$C_2\_v_1$	$C_3\_v_2$	$C_4\_v_1$	$C_5\_v_2$
11	$C_1\_v_3$	$C_2\_v_1$	$C_3\_v_2$	$C_4\_v_2$	$C_5\_v_1$
12	$C_1\_v_1$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_1$	$C_5\_v_1$
13	C1_v_3	$C_2\_v_2$	$C_3\_v_2$	$C_4\_v_3$	
14	$C_1\_v_1$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_3$	$C_5\_v_1$
15	$C_1\_v_3$	$C_2\_v_2$	$C_3\_v_2$	$C_4\_v_1$	$C_5\_v_1$
16	$C_1\_v_3$	$C_2\_v_2$	$C_3\_v_2$	$C_4\_v_3$	$C_5\_v_1$
17	$C_1\_v_1$	$C_2\_v_2$	$C_3\_v_2$	$C_4\_v_3$	$C_5\_v_2$
18	$C_1\_v_1$	$C_2\_v_2$	$C_3\_v_2$	$C_4\_v_2$	$C_5\_v_2$
19		$C_2\_v_1$	$C_3\_v_1$	$C_4\_v_2$	$C_5\_v_2$
20	$C_1\_v_2$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_2$	$C_5\_v_2$
21	$C_1\_v_2$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_3$	$C_5\_v_2$
22	$C_1\_v_2$	$C_2\_v_1$	C3_v_2	$C_4\_v_1$	$C_5\_v_2$
23	$C_1\_v_3$	$C_2\_v_2$	$C_3\_v_1$	$C_4\_v_2$	$C_5\_v_1$
24	$C_1\_v_2$	$C_2\_v_1$	$C_3\_v_2$	$C_4\_v_3$	$C_5\_v_2$

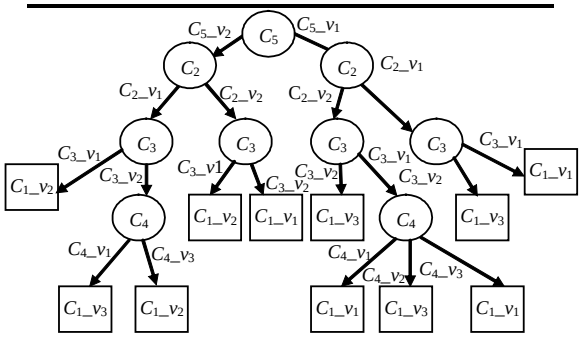

图1　ID3算法生成的决策树(以C_1为叶子节点)

决策树中椭圆表示属性，矩形表示叶子节点。根据表1中从根节点到叶节点的路径，可以得到如下的分类规则：(1) IF(C_5=$C_5\_v_1$ AND C_2=$C_2\_v_1$ AND C_3=$C_3\_v_1$) THEN (C_1=$C_1\_v_1$)；(2) IF (C_5=$C_5\_v_1$ AND C_2=$C_2\_v_1$ AND C_3=$C_3\_v_2$) THEN (C_1=$C_1\_v_3$)；(3) IF (C_5=$C_5\_v_1$ AND C_2=$C_2\_v_2$ AND C_3=$C_3\_v_1$ AND C_4=$C_4\_v_3$) THEN (C_1=$C_1\_v_1$)；(4) IF (C_5=$C_5\_v_1$ AND

C_2=C_2_v_2 AND C_3=C_3_v_1 AND C_4=C_4_v_1) THEN (C_1=C_1_v_1)；(5) IF (C_5=C_5_v_1 AND C_2=C_2_v_2 AND C_3=C_3_v_1 AND C_4=C_4_v_2) THEN (C_1=C_1_v_3)；(6) IF (C_5=C_5_v_1 AND C_2=C_2_v_2 AND C_3=C_3_v_2) THEN (C_1=C_1_v_3)；(7) IF (C_5=C_5_v_2 AND C_2=C_2_v_1 AND C_3=C_3_v_1) THEN (C_1=C_1_v_2)；(8) IF (C_5=C_5_v_2 AND C_2=C_2_v_1 AND C_3=C_3_v_2 AND C_4=C_4_v_1) THEN (C_1=C_1_v_3)；(9) IF (C_5=C_5_v_2 AND C_2=C_2_v_1 AND C_3=C_3_v_2 AND C_4=C_4_v_3) THEN (C_1=C_1_v_2)；(10) IF (C_5=C_5_v_2 AND C_2=C_2_v_2 AND C_3=C_3_v_2) THEN (C_1=C_1_v_1)；(11) IF (C_5=C_5_v_2 AND C_2=C_2_v_2 AND C_3=C_3_v_1) THEN (C_1=C_1_v_2)。

表2 填充属性C_1

索引	C_1	C_2	C_3	C_4	C_5
8	C_1_v_3	C_2_v_2	C_3_v_2	C_4_v_1	C_5_v_1
19	C_1_v_2	C_2_v_1	C_3_v_1	C_4_v_2	C_5_v_2

表3 填充属性C_5

索引	C_1	C_2	C_3	C_4	C_5
4	C_1_v_2	C_2_v_1	C_3_v_2	C_4_v_2	C_5_v_2
13	C_1_v_3	C_2_v_2	C_3_v_2	C_4_v_3	C_5_v_1

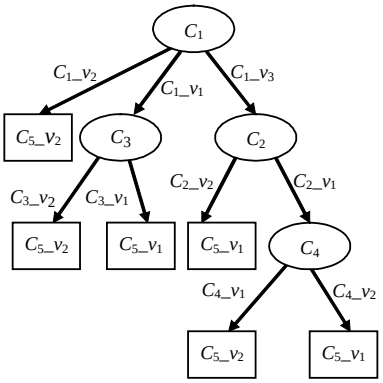

图2 ID3算法生成的决策树(以C_5为叶子节点)

根据分类规则，可以预测出表1中第8个数据点中属性C_1的值为C_1_v_3，第19个数据点中属性C_1的值为C_1_v_2，最终填充后如表2所示。表1中还有一个不完整属性C_5，以C_5作为叶子节点可以得到如图2所示的决策树。根据图2也可得到对应的分类规则，从而可以预测出表1中第4个数据点中属性C_5的值为C_5_v_2，第13个数据点中属性C_5的值为C_5_v_1。

4 结 束 语

采样数据中往往存在一些不完整数据，而有的不完整数据有一定的可用性，不能轻易丢弃，因此，不完整数据的处理便成了数据挖掘中的一个重要研究课题。不完整数据总是和样本空间中其他完整数据存在一定的相似性，也就是通过完整数据，必定能够推导出满足样本空间所有数据的基本规律。基于此，本文提出了一种基于决策树的不完整数据的处理方法。在对不完整数据处理方法的比较、分析的基础上，采用一种有效的决策树方法对不完整数据处理，最后，以一个实例验证了该方法的效果。该方法扩展了决策树的应用，对数据挖掘的研究有一定的促进作用。

参 考 文 献

[1] LEE N C. Improving data quality: development and evaluation of error detection methods[D]. Taiwan: National Sun Yat-Sen University,2002

[2] HAN J, Kamber M. Data mining: Concepts and techniques[M]. San Francisco: Morgan Kaufmann, 2000.

[3] 张 颖. 数据采掘的研究与应用[D]. 北京:中国科学院计算技术研究所, 1998.

[4] 毛国君, 段立娟, 王 实, 等. 数据挖掘原理与算法[M]. 北京: 清华大学出版社, 2005.

[5] 陈 伟, 丁秋林. 数据清理中不完整数据的清理方法[J]. 微型机与应用, 2005, 24(2): 44-55.

编 辑 熊思亮