决策树中数字型连续属性的语义化研究①

Numerical Continuous-Valued Attributes Semanteme in Decision Tree

周笑天 (山东省气象信息中心 山东 济南 250031)

摘 要: 针对目前大多数决策树挖掘中处理连续型属性方法时不考虑语义信息的问题，指出了研究数字型连续属性的语义化问题的必要性和可行性，进而提出了决策树中数字型连续属性的语义化方法，最后结合实例对该方法进行了验证。

关键词: 数据挖掘 决策树 数字型连续属性 语义

数据挖掘(Data Mining，简称 DM)从数据库中寻找数据中的模式和数据间潜在的关联，它是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[1]。

而决策树数据挖掘（Decision Tree Data Mining）是数据挖掘中分类算法的一种，决策树叶可称为判定树，它是运用于分类的一种树结构。

在决策树方法中，首先根据训练集数据形成决策树，如果该树不能对所有对象给出正确的分类，那么选择一些例外加入到训练集数居中，重复该过程一直到形成正确的决策集，因此决策树是代表着决策集的树形结构[2]。

1 决策树挖掘中数字型连续属性的语义问题

在线有的决策树挖掘系统中，决策树所能处理的样本数据集都是不包含语义信息的，特别是对于数字型连续属性的处理，人们大多是通过一些固定的方法进行离散化后再运用决策树相关算法进行挖掘。目前对连续属性的值进行离散化划分具有多种方法，现有实验已经证明所有可能划分状态的最优离散化方法是一种 N P - hard 问题. 对连续属性离散化的方法目前有三种分类: 其一,有监督的离散化和无监督的离散化[3]; 其二全局离散化与局部离散化; 其三,静态离散化与动态离散化[4]。

在将连续属性离散化时，人们只是将连续属性的取值离散为某个具体数据类，使得在进行数据挖掘前连续属性的取值就被限制为一系列的数据段或相关的几个离散值，并没有考虑到相关的数字所对应的语义，随着决策树算法的发展和应用，决策树挖掘中数字型连续属性的语义化问题成为人们需要解决的一个技术问题。

在现实世界里，每个实体都可以用多个特征来描述，每个特征又都有自身的意义和相关具体值来量化。如实体是某天的天气，它有如下的特征：Outlook、Temperature、Humidity、Windy，则某天的天气可以描述为以下一个组合：（Outlook： sunny；Temperature：27℃；Humidity：50%；Windy：3 grade）。我们考虑其中的某一数字型连续属性，如 Temperature，假设给出该属性的一组训练集，如：（27，28，29，31，25，32，29，30，31，29），对于上述每个训练数据，在进行数据挖掘和处理时它们是相互孤立，毫无联系的个体，同时也不具备语义相关性，但结合它所属的属性名称为 Temperature，由于 Temperature 作为客观属性实体具有相应的语义分类，如：炎热，寒冷，凉爽。因此如果把数字型连续属性进行相关的语义映射，则数字就可具有语义性：对于数字 31，可以映射为天气较炎热；而对于数字 25，则可以映射为天气较凉爽。

数字型连续属性所隐含的语义问题在数据库中无

① 收稿时间:2008-08-16

法体现，在进行传统的决策树挖掘时对数字型连续属性挖掘时也没有考虑它的语义，因此，本文提出基于将数字型连续属性语义化的思想，并探索数字型连续属性语义化的方法，尝试将数字型连续属性语义化后再引入决策树挖掘中，从而解决传统的决策树挖掘缺乏对数字型连续属性的语义处理问题，同时也简化决策树的知识库并简化决策树的建立过程。

2 数字型连续属性的语义化方法

2.1 相关概念及定义

定义 1.对于概念C，若存在概念$C_1,C_2,\cdots,C_n$，满足对$\forall C_i \subset C \quad (i \in 1,2,\cdots n)$且，$C_1 \cap C_2 \cap \cdots \cap C_n \equiv \Phi$则称概念$C_1,C_2,\cdots,C_n$是概念$C$的语义子概念，称集合$\{C_1,C_2,\cdots,C_n\}$为概念$C$的一个语义概念类。

定义 2.对于概念C，若存在一个语义概念类$\{C_1,C_2,\cdots,C_n\}$，称$R_c=\{C_1,C_2,\cdots,C_n\}$为概念$C$的一个语义类视图。

定义 3.对于概念C，若存在$R_c=\{C_1,C_2,\cdots,C_n\}$的一个语义类视图，则可将概念$C$表示为一棵由$C$为根、其子概念$C_1,C_2,\cdots,C_n$为叶子结点的语义概念树。

由定义 3 可以推知，若概念C的子概念$C_1,C_2,\cdots,C_n$也存在子概念，则概念C的语义概念树的深度$d(C)>2$，从而可以推出$d(C)=\max\{d(C_i)\},i\in 1,2,\cdots,n$。

2.2 语义化算法

对于数字型连续属性的语义化问题，在进行语义化时，首先我们要考虑该属性的语义元素，分析该属性根据语义成分可以做出的相关划分，例如针对年龄这一连续属性，它的语义元素为：（attribute|属性,age|年龄,animate|生物）。结合现实生活，我们可以考虑年龄的一个语义概念类为:{少年，青年，中年，老年}；其次还要兼顾训练集中所给出的具体训练数值及其分布规律，因为我们的目的是将训练集中数据映射为语义概念类中的某个概念，而数据分布情况会影响数据与语义子概念的对应。所以我们在进行数字型连续属性语义化时要兼顾这两者。

针对数字型连续属性的语义化问题，我们提出通过构建数字型连续属性的概念树的方法来解决，同时给出构建数字型连续属性的概念树的方法。

数字型连续属性语义化的基本思想为：

将属性取值的数据放入一张二维表中，利用该属性的语义类视图$R=\{C_i\},i\in 1,2,\cdots,r$以及 k-means 聚类算法（聚类个数 k 为关系视图 R 中的某一层次的节点数），将数据划分成相应的集合以对应相应关系视图中的概念。

数字型连续属性语义化算法如下：

输入：某数字型连续属性的一组训练数据数组 Train[N]，该属性的语义类视图 R

输出：某数字型连续属性取值集分类后所对应的语义概念树

Step 1.选取语义类视图 R 中的某一深度层，计算该层结点数 n，令 k=n；

Step 2.循环 Step 3 到 Step 4，直到每个聚类不再发生变化后转 Step 5；

Step 3.根据每个聚类对象的均值，计算每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分；

Step 4.重新计算每个（有变化）聚类的均值；

Step 5.将 k 个聚类与相应的语义类视图 R 中的概念相对应；

Step 6.Train[N]分类后所对应的语义概念树表示。

通过上述方法，我们可以便可以将数字型连续属性语义化，并给出该属性相应的语义概念树的直观表示。

对数字型连续属性进行语义化后，可以把数据抽象至不同的语义概念层次再进行相关的决策树挖掘，这会使得挖掘比在原始层更有优势，也能使挖掘的数据变得简单且在某种程度上还能表达出相应的语义的信息。此外，先将数字型连续属性语义化后再进行挖掘得到的挖掘结果决策树，在沿着由根到树叶节点遍历产生分类规则时也可以表达语义信息。

2.3 数字型连续属义语义化决策树算法的调整

将数字型连续属性语义化后，决策树 ID3[5]算法中的相关概念不需调整，给定的样本集合 T 分类的期望信息仍为：

$$\inf o(T)=I(P), \tag{1}$$

其中$C_1,C_2,\cdots,C_k$为集合 T 根据类别属性的值所分成的相互独立的类；P是的概率分布$(C_1,C_2,\cdots,C_k)$，即：

$$P=(|C_1|/|T|,|C_2|/|T|,\cdots,|C_k|/|T|) \tag{2}$$

但在计算由数字型连续属性将训练样本进行划分信息量时所得的期望信息则会不同：

$$E(A)=\sum_{i=1}^{v}\frac{S_{ij}+\cdots+S_{mj}}{S}I(S_{ij},\cdots,S_{mj}), \tag{3}$$

公式(2)中$(S_{ij}+\cdots+S_{mj})/S$是第j个子集的权，它是由所有子集$\{S_1,S_2,\cdots,S_v\}$中取值为a_j中的

样本个数除以 S 中的样本总数。数字型连续属性语义化后由于数据进行了概念爬升，相应的数据表所划分的第 j 个子集会比原数据表涵盖的范围广，相应的表数据元素也相应增加，计算 $E(A)$ 时 $S_{ij}+\cdots+S_{mj}$ 会改变，因此所需计算的 $I(S_{ij}+\cdots+S_{mj})$ 也会改变，但 S 作为数据库的训练样本总数不会改变。

信息增益 $Gain(X,T)$ 仍为：

$$Gain(A)=\inf\ o(T)-E(A) \qquad (4)$$

3　应用实例

人们日常生活中的各项活动都在一定程度上受到天气情况的影响，天气情况不只包含人们通常所理解的天气的阴、晴状况，温度、空气湿度以及风力情况都是会影响到人们生活和各项活动的气象因素。

我们将数字型连续属性语义化理论应用到一个气象系统模型实例当中，该实例分析的是不同天气、温度、湿度和风力的情况下是否适合进行户外运动的数据模型。给定该模型的训练数据集如表 1 所示：

表 1　天气情况对户外运动影响的训练实例集

序号	天气	温度 (℃)	湿度 (%)	风力 (级)	户外运动是否合适
1	晴	37	56	1	不合适
2	晴	36	55	4	不合适
3	多云	37	57	1	合适
4	有雨	31	59	2	合适
5	有雨	25	46	2	合适
6	有雨	26	48	5	不合适
7	多云	24	49	6	合适
8	晴	30	55	2	不合适
9	晴	25	49	3	合适
10	有雨	32	47	1	合适
11	晴	31	49	5	合适
12	多云	30	60	6	合适
13	多云	36	46	2	合适
14	有雨	26	61	5	不合适

对每个值 $A_j(j=1,\cdots,m)$，将所有记录划分成两部分，即：$\le A_j$ 和 $\ge A_j$，针对每个划分分别计算信息增益，选取最大信息增益的值进行划分。该方法显然会增加运算量和计算的复杂度，特别是对于包含多个连续型属性的上述训练实例，使用该方法更会造成复杂和冗余的重复的计算。

现在利用上面提出的数字型连续属性语义化的方法来将气候训练集中的各个数字型连续属性分别进行语义化：对于属性“温度”，首先将“温度”的训练数据 temperature[14]放入一张二维表中，可得图 1(纵坐标数据统一定为 10)。

图 1　属性“温度”数据二维表

给定属性“温度”的语义概念类 R 温度={炎热，凉爽，寒}和相应的属性的概念树(如图 2 所示)：

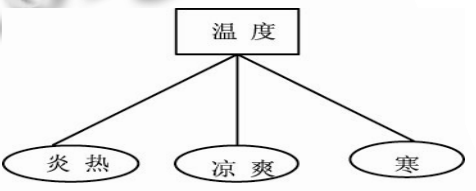

图 2　属性“温度”概念树

选取该语义类视图 R 的深度 h=2，计算该层结点 n=3，则聚类个数 k=3。

按照算法，属性“温度”的训练数据集 temperature[14]聚类后的结果为：(37，36)，(30，31)，(24，25，26)。将聚类结果与语义类视图 R 中的各个概念相对应，即为：(37，36)==炎热，(30，31)==凉爽，(24，25，26)==寒。

将 temperature[14]语义化分类后所对应的语义概念树表示为图 3。

类似的处理方式，给定属性“湿度”的语义概念类 R 湿度={高湿度，中湿度，低湿度}，属性“湿度”的训练数据集 humidity[14]聚类后的结果为：(60，61，59)，(55，56，57)，(48，49，47，46)，对应为：(60，61，59)==高湿度，(55，56，57)==中湿度，(48，49，47，46)==低湿度。

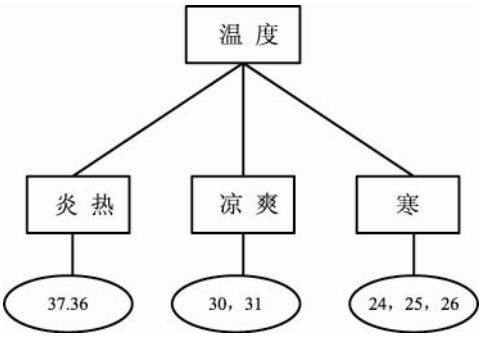

图 3　属性“温度”语义化后的概念树

将 humidity [14]语义化分类后所对应的语义概念树表示为图 4：

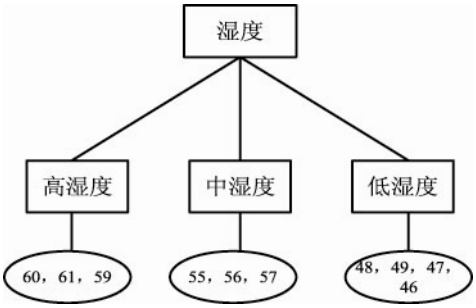

图 4　属性“湿度”语义化后的概念树

给定属性“风度”的语义概念类 R 风力={有风，无风}，属性“风力”的训练数据集 windy[14]聚类后的结果为：(1，2，3)，(4，5，6)，可对应为：(1，2，3)==无风，(4，5，6)==有风。

将 windy[14]语义化分类后所对应的语义概念树表示为图 5：

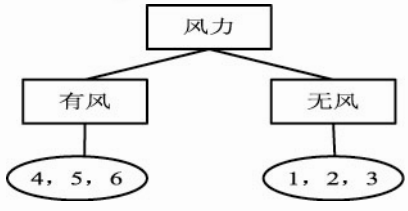

图 5　属性“风力”语义化后的概念树

利用所建立的各个数字型连续属性的相关语义概念树，对各数字型连续属性“温度”、“湿度”、“风力”进行语义化后，重新应用决策树挖掘算法(ID3 算法)挖掘所得结果如图 6：

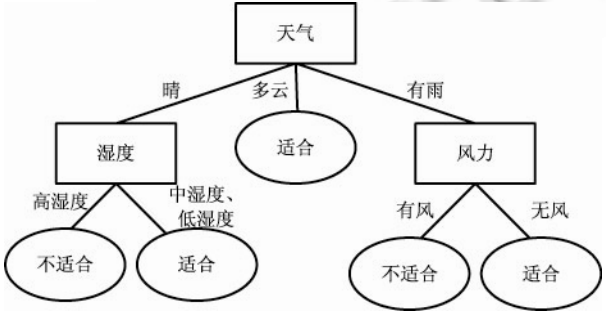

图 6　数字型连续属性语义化后的挖掘结果

由实例结果我们可以看出，将数字型连续属性语义化后再进行决策树挖掘结果更简洁也更能表现连续型属性的语义信息。我们可以通过挖掘结果得到相应的语义信息：天气晴朗且中低湿度的天气、多云的天气以及有雨且无风的天气都适合户外运动。

4　数字型连续属性语义化问题的讨论

分析我们提出数字型连续属性的语义化方法，可以看出，数字型连续属性的语义化结果主要有以下两个方面决定：该属性的语义类视图 R，以及语义化时所选取的深度 h。此两者共同决定了数字型连续属性语义化后所对应的概念类。

如果给定数字型连续属性的语义概念类不同，则挖掘结果会表现出不同的语义，例如，若我们最初给定属性“温度”的语义概念类为 R 温度={热、冷}，R 湿度={高湿度、低湿度}，R 风力={轻风，强风，大风}，则挖掘结果会表现出上述各个语义类组合形成的天气对人们能否参加户外活动的影响情况。若上述试验中的属性“风力”的语义类视图 R 风力={无风，有风}又有语义子类视图，如：R 有风={轻风，强风}，此时则可选取 h=3 将“风力”进行语义化，则“风力”的各个训练数据聚类后将对应如下的语义概念：无风，轻风，强风。

5　总结

将数字型连续属性进行语义离散化后用在决策树挖掘中，不但可以提高挖掘系统的知识表示能力、在一定程度上改善决策树挖掘知识库记录重复或语义模糊等问题，还能通过对数字型连续属性字段概念的语义化对与挖掘相关的各个属性进行深层的知识表示，从而在一定程度上使基于语义的决策树挖掘和预测成为可能。

参考文献

1 张云涛,龚玲.数据库挖掘原理与技术.北京:电子工业出版社, 2004.
2 邵峰晶,于忠清.数据挖掘原理与算法.北京:中国水利水电出版社, 2003.
3 Tay FEH, Shen Lx. A modified chi2 algorithm for discretization. IEEE Transactions on Knowledge and Data Engineering, 2002, 14 (3): 666－670.
4 Stefanowski J. Handling continuous attributes in discovery of strong decision rules. Rough Sets and Current Trends in Computing, 1998. 394－401.
5 Quinlan JR. Induction of decision trees. Machine Learning, 1986,1(1): 81－106.