基于效用的结构语法的属性学习

杨祥茂　黄　涛　周启海

（西南财经大学信息技术应用研究所　成都 610074）（西南财经大学经济信息工程学院　成都 610074）

摘　要　无论是在机器学习还是在软件设计中，对问题的分析都是假定对概念属性已知的条件下展开的。本文采用假设对象是在结构语法的基础上，通过确定对象的领域和效用，用领域的条件和状态及其边际效用选择出学习的类。在一个对象类中，属性的选择学习用迭代前向逐步插入、迭代向后删除算法。对决策树的学习，设计了决策树遗传归纳算法学习。对于新构造属性，用属性效用的投影、积、差、值、最值、级数算法构造边际效用法。对不同的对象进行不同的表达式学习，并命名新属性，进而进行属性的数据类型有界限定，从而得到新属性。

关键词　属性，效用，归纳学习，属性学习

Attributer Learning Based on Structure Grammar and Utility

YANG Xiang-mao　HUANG Tao　ZHOU Qi-hai

(Research Institute of Information Technology Application, Southwestern University of Finance and Economics, Chengdu 610074, China)

(School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu 610074, China)

Abstract　For machine leaning and software design, dealing with and analyzing concept is based on the attribute of it. This paper hypothesizes the object is a structure grammar, which is determined by the domain and utility of the object, and based on the structure grammar, and chose out for learning by the terms and conditions in its domain and marginal utility. In the structure of a object, the learning chooser of attributes needs to use the Prior to Gradually Insert Iterative Method, Delete Iterative Algorithm Backwards Method and the Decision Tree learning, then design the Decision Tree Algorithm Genetic Summarized Method. For the attributes newly constructed, it learns different expressions to different objects by constructing Marginal Utility Method by projection, multiplication, division, subtraction, extreme value, maximum and minimum, Series method of Attributes' Utility, and gives a name to the new attributes. Then it restricts the type of attributes' data to get the new attributes.

Keywords　Attribute, Utility, Inductive learning, Attributive learning

1　引言

在机器学习中，概念学习是把实例表示为确定属性的预定义假设空间中从特殊的训练样例中学习一般概念；决策树学习把实例表示成为树型结构，树上的结点就是对实例的某个属性(attribute)的测试，其后继分支对应着该属性的一个可能取值；人工神经网络的学习把学习过程表示为一系列相互连接的神经元构成，每个神经元经过一定数量的输入触发并产生单一输出的学习过程；基于统计的学习是建立在容量确定的情况下，根据经验用统计估计产生信号模型的方法；基于实例的学习预先建立简单的训练样例，当遇到新实例时触发原有的训练样例而产生新的查询分类，并根据新的分类产生一个目标函数给实例；基于规则的分析学习根据领域理论用逻辑分析方法分析并改进目标概念。所有这些学习都是基于已知确定属性的学习，如果能够对属性进行学习，则有利于问题对象的机器自动学习。

在软件设计中，对数据的描述与计算模型是根据数据类型而展开，其数据类型就是不同数据对象属性及其取值的集合，就是在抽象不同数据对象的属性并确定域取值范围。例如，操作系统中的进程就是抽象进程的属性(主要包括进程标识属性、进程状态属性、进程调度属性、进程通信属性、进程权限属性、进程资源使用与所有权属性等)，并且根据这些属性确定相应操作运算与管理运算；DBMS的系统设计中，设计就是抽象数据表中字段属性及其数据表之间的属性，从而建立起DBMS的设计，就是说通过抽象数据表中的字段属性(主要包括字段名、字段类型、字段宽度等)来设计数据表，在此基础上建立相应的存储结构，同时抽象数据表之间的属性来设计数据库；应用系统软件的设计，就是定义系统中概念的静态结构和动态行为，静态结构就是定义系统中对象的属性和操作及其对象相互之间的关系，动态行为就是定义对象的时间特性和对象为完成目标而相互进行通信的机制。当前，无论是在机器学习还是在软件设计中，人们都是假设已经确定了属性而展开的。这样，如何用机器学习的方法确定所论对象的有效属性就成为应予研究的内容，使得程序设计从手工劳动设计到面向领域的软件系统集成推进，从而有利于降低软件开发中的成本，提高软件开发的效率，以达到用软件来自动规范软件开发过程，提高软件系统可靠性之目的。

为了应对软件系统中属性问题的抽象过程与描述的复杂性，本文力图假设在结构语法的基础上，通过确定对象的领域和效用，用领域的条件和状态及其边际效用选择出学习的类。

杨祥茂　副教授，主要研究方向为计算机应用等；黄　涛　讲师，主要研究方向为计算机应用；周启海　教授，博(硕)士生导师，主要研究方向为计算几何、算法研究与应用、财经计算、同构化信息处理等。

在一个对象类中，属性的选择学习用迭代前向逐步插入、迭代向后删除算法、对决策树的学习，设计了决策树遗传归纳算法学习。对于新构造属性，用属性效用的投影、积、差、极值、最值、级数算法构造边际效用法对不同的对象进行不同的表达式学习，并命名新属性，进而进行属性的数据类型有界限定，从而得到新属性。

2 基本概念

2.1 属性概念

任何对象都是由其属性来表示的，其属性的组合反映了对象的内涵。

定义 1(属性) 属性就是论域中的元素与命题的一种有效映射，元素就是论域中真值的一种映射。

定义 2(命题) 命题就是对一个结构体(construction)描述的可判断的归类，一个属(属性值)就是论域中每一个元素同一个命题的一种关系(association)。

据此，属性就是论域 X 到命题 α 的函数，其形式化描述为

$$\text{attributer}(X,\alpha)::= X\odot\alpha$$

其中，attributer 就是属性函数名称，$\odot$ 就是命题为真的语义运算符，论域 X 就是对象数据的集合，α 为相映的命题。

性质 1(简单属性) 就是不可再分的属性。如果一个属性是简单属性，则它的任意组合与连接运算都不是简单属性，即不存在另一个属性使得同一论域到相同命题的函数相同。

性质 2(合属性) 就是简单属性的结构组合体。对象通过属性的组合体来实现，它可以是对象的成分、事实、时间事件、空间事件等结构组合体，通过结构组合体构成新对象的属性集合。

性质 3(属性蕴涵) 对象集合 G 中，属性 M 有 $A,B\subseteq M$，每个具有 A 中属性的对象就一定具有 B 中的属性，即属性蕴涵 $A\rightarrow B$。

性质 4(属性依赖) 对象集合 G 中，属性 $A,B\subseteq M$，对每个对象 $g\in G$，B 函数依赖 A，当且仅当存在一个函数映射 f，使得 $f:(m(g)\mid m\in A)=(n(g)\mid n\in B)$ 对所有的 $g\in G$ 成立。

在计算机的语言描述中，数据属性的描述即数据类型，字符的属性就是关于字符模型及其建立在模型上的操作的集合所构成的数据类型，数值属性就是关于整型数据和实型数据的抽象数学模型基础上的计算，结构数据的属性就是用集合结构、线性结构、树型结构、图结构的抽象数据类型模型来描述。在计算机模拟系统中，用于模拟人类肢体语言的词数据属性用三元组结构的数据模型来描述，即形数据(例如手形数据和肢体形数据)、方向数据、位置数据的结构模型。

2.2 结构语法的描述

按照 20 世纪 70 年代的布列斯南(J. Bresnan)和卡普兰(R. Kaplan)提出的 LFG 文法，它包括两个语法层次结构，即成分结构(Consituent Structure)和功能结构(Function Structure)。成分结构称为 C-结构，功能结构称为 F-结构，两者都表示了功能信息，这些功能都是用“属性-值”的有序对来表示，其功能结构的形式定义为：

(1)它是有序对的集合，每个有序对包含一个属性和该属性的值。

(2)语法功能的名称或者特征名称就是属性。

(3)属性值主要有：简单符号、语义形式、子结构。

由此，具有结构语法的语句就可以表示为符号的前缀和后缀序列，而其相应符号文本则可对应表示为序列文件，且可记为 $\{\alpha_1\beta_1\alpha_2\beta_2\alpha_3\beta_3\alpha_4\beta_4\cdots\cdots\}$[4]。

2.3 效用与边际效用

效用是对象属性子集的取值。

建立效用函数的假设前提：①有两个可能预期属性 A，B，则 $A=B$ 或者 $A\neq B$。②预期属性 A,B,C，若 $A>B,B>C$，则 $A>C$。③任意属性的概率组合为属性集的并、交、补、幂集运算。

定义 3(效用函数 u) u 为预期属性的整数，对任意个体属性 x，有效用函数 u，$u=u(x)$。

设 A,B 分别为属性，并且 $A\neq B$；则有如下性质。

性质 5(等价) A 包含 B，等价于 $u(A)>u(B)$。

性质 6(线性选择关系) 若 $0\leqslant p\leqslant 1$，则 $u[pA+(1-p)B]=\{u(A)\mid u(B)\}$，即组合后的效用值要么是 $U(A)$，要么是 $U(B)$。

定理 1(边际效用，$\frac{\partial u}{\partial x}$) 当 x 递归包含 y，变化的层数 Δx，其概念不变。Δx 变化所产生的效用为 ΔU，即自变量最后一个单位的变化所带来的因变量的变化。

在成分结构的语法规则下，结构中的每个属性的效用值初值为 1。如果某个属性还可以进行分层，则效用值依次加 1。就是说，属性可以形成概念分层，用基数值来描述。一个属性的不同值用层数的多少来确定属性的基数值，例如 street<city<province_or_state< country。也可以用集合包含来描述效用值，例如{chengdu, nanchong}包含在 sichuan。

3 基于效用的结构语法的属性学习模型

基于效用的结构语法的属性学习模型是一个 4 元组的输入输出系统模型，它包括输入结构语法的输入序列 I、输出对应的属性集 O、结构语法约束集 T、模型中的基于输入序列的学习关系集 R，即基于效用的结构语法的属性学习模型为 $M=(I,O,T,R)$。

换言之，在输入序列的语句子集 $\alpha_i\beta_i$，通过学习关系集 R 的算法识别，如果后缀 β_i 在知识库中有相同的知识，通过效用判断其是简单属性还是合属性。如果其效用值为 1，则认为是简单属性，非 1 则是合属性。对于合属性进而反向学习其属性。如果知识库中没有相同的后缀，则假设是一个简单属性，并新建一个属性，命名为一个新的“属性”，加入到属性库中。

3.1 属性子集选择

属性子集的选择，通过删除不相关或者冗余的属性获得最小属性集，使得新的属性保持原来的分布。设属性集 A，属性效用值集合<，在二元组 $(A,<)$ 的属性选择学习操作主要有：

(1)迭代前向逐步插入归约。其算法思想可简要描述如下：

算法 1 迭代前向逐步插入算法

输入：初始属性集合$\{A_1,A_2,\cdots,A_i,\cdots,A_n\}$
输出：归约属性集合 A
算法开始：
①清空初始集合 A；
②依次对属性集合$\{A_1A_2,\cdots,A_i,\cdots,A_n\}$中的每个元素进行如下处理；
③确定 Ai 的边际效用；
④如果 A_i 的边际效用等于 0，则不插入 A_i；否则，插入 A_i 到属性集合 A 中；
⑤输出集合 A；
算法结束。

(2)迭代向后逐步删除归约。其算法思想,基本类同于算法1,故略。

(3)决策树遗传归纳。在每个节点选取"最好"的属性,并将其分类。

算法2 决策树遗传归纳算法

输入:初始属性集合$\{A_1A_2\cdots A_i\cdots A_n\}$
输出:归约属性集合 A
算法开始:
①置 k=0,随机产生属性效用值的初始种群 $\vec{X}(0)$:
$\vec{X}(0)=(X_1(0),\cdots,X_N(0))\in\{A_1A_2\cdots A_i\cdots A_n\}$;
②独立地从当前种群中选取 N-1 个母体;
③独立地从所选取的 N-1 个母体进行边际效用值杂交,得到 N-1 个中间母体;
④独立地对杂交后的 N-1 个个体进行 $\frac{U'(x_i)}{\sum_{i=1}^{N}U'(x_i)}$ 变异,得到 k+1 代种群的前 N-1 个个体:
$X_1(k+1),\cdots,X_{N-1}(k-1)$;
⑤计算 $i=\arg\max\{U'(X_j(k))\}$,令 $X_N(k+1)=X_i$,则停止;
⑥若不满足,则 k=k+1 并返回②;
算法结束。

3.2 新属性的学习

新属性的学习主要包括属性初始命名学习、属性归纳与分析命名。

(1)属性初始命名学习

若语句结构体的语义值为真,从取值集合中按照多值对应的原则。如果属性库中无相应的属性记录,则确定其属性并命名其名称,按照概念的成分、性质与事实来确定属性。例如,假设服装类库中没有颜色的属性,但是在叙述中存在说"服装是雪色",就是说雪的属性值是白色,其属性名称命名为"颜色",并且设定该属性的效用初值为1。

(2)属性的归纳与分析命名学习

通过数据为属性的训练样例进行属性泛化,利用一个边际效用属性函数,得到新属性保持原有的数据一致性。在属性归约中,常对已知数据进行小波变换、主成分分析、线性变换、对数变换、桶变换(直方图等)、聚类(距离变换)等,基本算逻辑运算的函数变换,本算法中我们用属性效用的投影、积、差、极值、最值、级数边际效用法对不同的对象进行不同的表达式学习,进而命名新属性,进行属性的数据类型有界限定,从而得到新属性。

算法3 属性效用的边际效用涣化算法

输入:初始属性集合$\{A_1A_2\cdots A_i\cdots A_n\}$
输出:新命名属性
算法开始:
①FOR i=1 TO n;
② FOR j=1 TO i;
③ 利用所有 A_j 的边际效用值,施行投影、积、差、极值、最值、级数变换;
④ 如果变换后保持原来的数据一致性,则命名新的属性;
⑤ ENDFOR j;
⑥ ENDFOR i;
⑦构造新属性的数据类型,得到新的属性;
算法结束。

据此,便可设计基于效用的结构语法属性学习模型系统。例如,在某服装厂服装属性学习系统实验设计中,便可采用服装作为所论对象来进行其概念的属性学习,使其输入文本是一段该服装厂关于客户需求的说明,通过服装的客户需求,分解单词,确定所有的 β_i 和 α_i,并对每个 α_i,β_i 进行属性学习。从而通过对属性的自动学习,来获得其服装客户的各属性(包括姓名、性别、电话、接件日期、取件日期、上体长、手臂长、胸围、颈围、腰围、臀围、肩宽、胸宽、背宽、前腰节高、后腰节高、总体高、身高、下体长)。

结束语 在属性的学习中,把属性当成数据进行分析和归纳,目前有许多问题尚待解决。例如,学习环境的描述及其设计问题;对于已知环境为随机事件并且已知其分布,可采用基于统计的机器学习法;结构语法文本分析;文本单词分解;等等。但由于其复杂性,本文仅对结构语法的文本进行了初步分析和一定研究,并假定文本是结构语法描述的,且为简单的结构语法,故仅为引玉之砖。进一步的研究,至少应有:如果文本不是简单的结构语法描述,而是复杂的结构语法描述,则需要对文本的单词用复杂的结构语法进行有效分解,才能得到 α_i 和 β_i;学习的表达式构造方法绝不止用属性的效用和边际效用来构造,故相关算法也应进行深入的分析、研究和设计;数据的数据类型和范围尚需更详细的数据类型描述工具(显然,如果数据类型的描述分类更小,则理当为属性的自动学习找到更有效的方法。比如,顾客数据包括年龄和年薪属性,要根据实际情况进行属性的规范化处理,年薪的取值范围可能比年龄的取值范围要大得多)。此外,在实际应用系统的设计中,还可用中间件、组件等技术思想,对不同类施行不同的属性自动学习。

参考文献

[1] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. Second Edtion. 机械工业出版社,2006
[2] 杨祥茂,等. 基于网络资源消费者模型的调度策略. 计算机科学,2003,30(9):105-106
[3] 杨祥茂,等. 基于均衡模型的计算机资源分配. 计算机科学,2005,32(4):171-172
[4] 刘颖. 计算机语言学. 清华大学出版社,2002
[5] Michell T M. 机器学习. 机械工业出版社,2003

(上接第236页)

[2] Wulf W A, McKee S A. Hitting the Memory Wall: Implications of the Obvious. Computer Architecture News, 1995, 23(1): 20-24
[3] Bordawekar R, Choudhary A, Kennedy K, et al. A Model and Compilation Strategy for Out-of-core Data Parallel Programs// Proceedings of ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. ACM Press, 1995: 1-10
[4] Golumbic M C. Algorithmic Graph Theory and Perfect Graphs. A-nnals of Discrete Mathematics, 2004
[5] Fabri J. Automatic Storage Optimization// SIGPLAN '79: Proceedings of the SIGPLAN Symposium on Compiler construction. 1979: 83-91
[6] Johnson M R. Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., 1979
[7] Kierstead H A. The Linearity of First-fit Coloring of Interval Graphs. SIAM J. Discrete Math 1, 1988: 526-530
[8] Confessore G, Dell'Olmo P, Giordani S. An approximation result for the interval coloring problem on claw-free chordal graphs. Discrete Appl. Math., 2002, 120(1-3): 73-90
[9] Pemmaraju S V, Penumatcha S, Raman R. Approximating interval coloring and max-coloring in chordal graphs. J. Exp. Algorithmics, 2005, 10: 2-8
[10] West D B. 图论导引. 原书第2版. 李建中,骆吉洲,译. 机械工业出版社,2006
[11] Andersson C. Register Allocation by Optimal Graph Coloring// CC'03: Proceedings of the 12th International Conference on Compiler Construction. Springer-Verlag, 2003
[12] Li L, Nguyen Q H, Xue J. Scratchpad Allocation for Data Aggregates in Supperperfect Graphs. ACM SIGPLAN NOT., 2007, 42(7): 207-216
[13] Tanenbaum A S, Woodhull A S. 操作系统:设计与实现(第二版). 王鹏,尤晋元,朱鹏,熬青云,译. 电子工业出版社,1998