统计 信息 递归 机器

25-28

计算机科学 Vol.20 No.1 1993

统计信息、递归和机器发现*)

李爱中 (吉林大学计算机科学系 130023长春)

F222.1

摘 要

A new approach to machine discovery is presented in this paper. This approach uses statistical information as heuristics, applies primitive recursion as the basis for function representation and the structure control of the discovery algorithm. The discovery algorithm finds functions by recursive decompositions and transformations under the direction of statistical heuristics.

一、引言

1.1 问题

本文所要研究的中心问题是机器发现的数学描述，即：给出一个变元的集合$\{y, x_1, \cdots, x_n\}$和关于这些变元的一组数据$\{d_1, \cdots, d_n\}$，其中$d_i = (y_i, x_{1i}, \cdots, x_{ni})$满足条件：当$x_1 = x_{1i}, \cdots, x_n = x_{ni}$时，$y_i = y$。试图发现一个函数f使得$y = f(x_1, \cdots, x_n)$尽可能拟合数据。数据可以是定量的，也可以是定性的。但为了方便讨论，在本文中所讨论的变元取值限于非负的整数。

对此问题的研究涉及到两个领域：统计学中的数值分析和机器发现。二者截然不同。

1.2 统计学中的数值分析

在科学发展的进程中，归纳起着举足轻重的作用。科学定律都是通过归纳从大量事实中得到的。从发现的角度来说，演绎充其量是把归纳的结果特化为一些简单情形。传统上关于科学定律的归纳的研究属于统计学中数值分析的范畴。

在统计学的数值分析的意义下，拟合意味着求解预定形式的函数中的参数。参数在求出值后是常量。近来，一些学者把数值分析方法应用于机器发现来估计参数[8]。这种途径在曲线拟合上表现得很优越，但在发现函数上陷入了僵局。如果函数的形式是未知的并且这个未知的函数形式恰是我们必须发现的，也就是说我们所要发现的不仅仅限于函数的参数，这时传统的基于统计学的数值分析方法陷入了困境。

1.3 机器发现

对于上述问题的另外一种解决途径是机器发现。机器发现以Langley和Simon等人的开创性工作和其著名的机器发现系统BACON为代表。近年来，有很多机器发现系统不断推出，如：ABACUS[2]，FAHRENHEIT[3]和COPER[4]等。

作者认为：上述发现系统的发现能力是有限的。其根源在于数学基础和启发式信息。上述系统均以传统的数学分析为数学理论基础，恰恰是数学分析的存在性限制了发现过程的计算机实现，从而导致了计算复杂性的限制和发现能力的局限。简单的基于直觉的启发式信息在搜索比较大的函数空间时，不能克服复杂性的障碍。由此，产生了机器发现的另外一种研究途径。这种途径以递归函数

*) 国家863基金和自然科学基金资助研究项目

理论为理论基础并且具有构造性的（constructive）特点[6,7]。递归函数理论作为计算机科学论理论基础，这种途径很有发展前景。但基于递归函数理论的机器发现算法具有一定的局限性。自底向上的发现算法虽然发现能力得到改善，但无法克服复杂性的障碍[7]。自顶向下的发现算法由于缺乏启发式信息和不能很好地综合处理原始递归式和合成两个算子，因而在发现能力和复杂性上受到了限制[6,7]。究其根源在于缺少对算子进行综合性评价的启发式信息以指导搜索过程。

在本文中，作者试图利用统计信息作为启发式信息对算子进行评价，给出了一个构造性的发现算法。

二、统计信息

统计学中关于样本数据的计算信息称为统计信息。统计信息可以有不同的形式或定义。本文主要以线性相关系数为基础[1]，来定义统计信息。

线性相关系数是关于变元间线性相关程度的一种测度。若我们有关于变元（y，x），那么线性相关系数r可以定义为：

$$r=\left[\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right]\Big/\left[\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}\right],$$

而且$-1\leqslant r\leqslant 1$。

如果$|r|=1$，那么存在两个常量a和b，满足$y_i=a+bx_i$（$i=1,\cdots,n$）。因此，我们可以用$|r|=1$作为启发式信息，而发现函数$y=a+bx$，其中a和b可以由已知数据，用最小二乘原理计算出来。

类似地，我们可以定义多元线性相关系数r如下：设

$$l_{ij}=\sum_{k=1}^{n}(x_{ik}-\overline{x_i})(x_{jk}-\overline{x_j})\qquad(i,j=1,\cdots,m),$$

$$l_{i0}=\sum_{k=1}^{n}(x_{ik}-\overline{x_i})(y_k-\overline{y})\ (i=1,\cdots,m),$$

$$l_{00}=\sum_{k=1}^{n}(y_k-\overline{y})^2$$

$$\overline{y}=(1/n)\sum_{k=1}^{n}y_k,$$

$$\overline{x_i}=(1/n)\sum_{k=1}^{n}x_{ik}\qquad(i=1,\cdots,m),$$

$$L=\begin{vmatrix}l_{11}\cdots l_{1m}\\ \cdots\cdots\\ l_{m1}\cdots l_{mm}\end{vmatrix}\quad L^{-1}=\begin{vmatrix}c_{11}\cdots c_{1m}\\ \cdots\cdots\\ c_{m1}\cdots c_{mm}\end{vmatrix}$$

$B=L^{-1}L_0$，其中：$B=(b_1,\cdots,b_m)^*$且$L_0=(l_{10},\cdots,l_{m0})^*$；那么

$$r=\sqrt{\sum_{i=1}^{m}l_{i0}b_i}\Big/\sqrt{l_{00}},\ \text{且}\ 0\leqslant r\leqslant 1。$$

为了统一线性相关系数的定义方式，在下面我们将使用$|r|$作为启发式信息。一般，我们事先置定$|r|$的下限R，$0\leqslant R\leqslant 1$；若$|r|\geqslant R$，那么我们就认为变元间具有线性关系，并以此作为启发式信息而导致以线性函数作为发现的结果。

三、原始递归式和分解

原始递归式是一种算子[3]。若函数A$(x_2,\cdots,x_m)$和B$(x_1,\cdots,x_m,x_{m+1})$为已知函数，通过原始递归式可以定义一个新的函数$f(x_1,\cdots,x_m)$如下：

$f(0,x_2,\cdots,x_m)=A(x_2,\cdots,x_m)$

$f(x_1+1,x_2,\cdots,x_m)=B(x_1,\cdots,x_n,f(x_1,\cdots,x_n))$

原始递归式的显著特点是在定义新函数的定义中使用了已知的和正在定义的函数本身，并且可以保证新函数是可以计算的。

反过来，若我们已知关于函数$f(x_1,\cdots,x_m)$的数据，那么我们可以从$f(x_1,\cdots,x_m)$的数据中，利用分解，得到关于新函数A和B的数据：

$A(x_2,\cdots,x_m)=f(0,x_2,\cdots,x_m)$

$B(x_1,\cdots,x_m,f(x_1,\cdots,x_m))=f(x_1+1,x_2,\cdots,x_m)$

递归式和分解是下面给出的发现算法

的一个重要发现手段。分解不但在时间复杂性上得到好处，也可以提高发现能力。从理论上讲，递归式所定义的函数超出了仅使用合成所定义的函数的范围。

四、基于分解和变换的发现算法

4.1 合成和变换

如果函数 $h(x_1, \cdots, x_k)$，$g_1(x_1, \cdots, x_m)$，…，$g_k(x_1, \cdots, x_m)$ 为已知函数，那么通过合成算子可以定义新函数 $f(x_1, \cdots, x_m)$ 如下：

$$f(x_1, \cdots, x_m) = h(g_1(x_1, \cdots, x_m), \cdots, g_k(x_1, \cdots, x_m))$$

但是，反过来，若我们知道函数 $f(x_1, \cdots, x_m)$ 的数据，我们很难从中分解出函数 g_1，…，g_k 和 h 的数据。因为 h 变化，$g_1, \cdots, g_k$ 也随之变化，而 h 是很难单独确定的。因此，在下面给出的发现算法中，我们对 h 作了一些限制，形成一个有限集 H，H 具有下列特性：若 $h(x) \in H$，则 $h^{-1}(x)$ 是一个可计算的函数。

基于上述讨论，我们定义了基于 H 的变换。若函数 $f(x_1, \cdots, x_m)$ 的数据为已知，$h_i \in H$，通过变换，我们得到一个新函数 $f^1(x_1, \cdots, x_m)$：

$$f^1(x_1, \cdots, x_m) = h_i^{-1}(f(x_1, \cdots, x_m))$$

这样的限制，虽然从理论上会限制发现的能力或范围；但是若 H 选择得足够大，则不会限制在一定范围内的应用。

4.2 发现算法

发现算法先搜索线性函数，然后搜索非线性函数。非线性函数是通过在线性函数基础上应用算子：分解和变换来实现的。整个搜索过程由统计信息作为启发式信息来加以控制。

发现算法：

参数： R，数据和待发现的函数 f

步骤1. 计算统计信息 r；若 $|r| \geqslant R$，则发现 $y = a_0 + a_1x_1 + \cdots + a_mx_m$，$a_0$，…，$a_m$ 由最小二乘法确定，结束；否则，进入步骤2；

步骤2. 分解 f 的数据，形成函数 A 和 B 的数据；计算 A 和 B 的统计信息 r_A 和 r_B；

步骤3. 对于所有 $h_i \in H$，对 f 的数据实施变换，形成函数 f_i 的数据，计算 f_i 的统计信息 r_{fi}；

步骤4. 计算 $E = \max\{\min\{r_A, r_B\}, r_{f_1}, \cdots, r_{f_{|H|}}\}$；若 $E = \min\{r_A, r_B\}$，则以 A 和 B 作为待发现的函数并利用其数据递归调用此算法，转步骤1；若 $E = r_{f_k}$，则以 f_k 为待发现函数并利用其数据递归调用此算法，转步骤1；〔若 $E = \min\{r_A, r_B\}$，则 f 由 A 和 B 利用原始递归式构成；若 $E = r_{f_k}$，则 $f(x_1, \cdots, x_n) = h_k(f_k(x_1, \cdots, x_m))$〕。

4.3 示例

为了简捷，我们给出了二个简单的例子。其中例1包含了分解，例2包含了变换，

例1 待发现函数 f(x) 的数据如表1，R = 1，$H = \phi_0$，

步骤1：$|r| < 1$，转步骤2；

步骤2：分解 f 的数据形成 A 和 B 的数据，A = 0，B 的数据如表2；转步骤1；

步骤1：A = 0（常数），停止；

步骤1：B(x, f(x)) = 2x + f(x) + 1，$|r| = 1$，停止；故此：

$f(0) = 0$

表1

f(x)	x
0	0
1	1
4	2
9	3
16	4
25	5

表2

f(x+1)	x	f(x)
1	0	0
4	1	1
9	2	4
16	3	9
25	4	16

$f(x+1) = 2x+f(x)+1$。

例2 待发现函数f(x)的数据如表3(略)，R=1,H={1/x}。

步骤1：|r|<1，转步骤2；

步骤2：$r_B<1$

步骤3：$r_{h_1}=1$；

步骤4：$E=r_h'$，f_1的数据如表4(略)；转步骤1，

步骤1：$f_1(x)=x/2$，因此$f(x)=h_1(f_1(x))=2/x$。

五、结论

在利用统计信息作为启发式信息和应用递归函数理论作为理论基础的前提下，作者在本文中提出的机器发现的新方法，通过分解和变换来发现函数。启发式信息使得发现算法更有效，分解提高了发现能力和效率。另外，此方法由于以递归函数理论为基础，具有结构性特点，并便于计算机实现。

参考文献

[1] Byrkit, Donald R. (1987), Statistics Today, A Comprehensive Introduction, The Benjamin/Cummings Publishing Company, Inc.

[2] Falkenhailer, B.G.et al. (1986), Intergrating Quantitative and Qualitative Discovery: The ABACUS System, Machine Learning, 1: 367—401

[3] Kleene, S.C. (1952), Introduction to Metamathematics, Van Nostrand

[4] Kokar, M.M. (1986), Determing Arguments of Invariant Functional Description, Machine Learning, 1: 403—422

[5] Langley, P., et al. (1987), Scientific Discovery: An Account of the Creative Process, Cambridge, MA: MIT Press

[6] 李爱中 (1991),模型发现和智能决策支持系统工具的研究，哈尔滨工业大学博士学位论文

[7] Lin,Xiaofeng and Ungar, Lyle(1989), Inventing Theoretical Terms in Inductive Learning of Functions——Search and Constructive Methods, In: Ras, Zbigniew W. (Ed), Methodologies for Intelligent Systems, 4, 132—141

[8] Lubinsky, David, et al. (1987), Data Analysis as Search, In: Phelps, Bob (ed.), Interactions in Artificial Intelligence and Statistical Methods, Unicom Seminars Ltd

[9] Zytkow, J.M.(1987), Combining Many Searches in the FAHRENHEIT Discovery System, In: Proceedings of the Fourth International Workshop on Machine Learning, Irvine, CA. 281—287

(上接57页)

参考文献

[1] L.G.Shapiro et all,Relational Matching,Apll.Opt.,26: 10(1987), 1845-1851

[2] T.Pavilids,Structural Pattern Recognition, Springer Verlag, 1977

[3] D.W.Tank and J.J.Hopfiled, Simple 'naural' optimization network, An A/D Converter, Signal Decision Circuit and a Linear Programming Circuit, IEEE Trans. Circuits & System, CAS-33:5 (1986), 533-541