优先归纳逻辑程序的极限行为

马世龙[1,2] ,眭跃飞[2] ,许 可[1]
(1. 北京航空航天大学 计算机学院,北京 100083;2. 中国科学院 计算技术研究所,北京 100080)

摘 要:虽然对归纳逻辑程序的极限行为至今并没有深入的研究,但是通常在分析正在执行的增量式或在线归纳学习算法时,必须考虑这种程序的极限行为. 某些归纳学习算法如果不考虑极限行为可能运行到最后会发生错误. 如果给定一个递增的例子集合序列,一个归纳逻辑程序会产生一个相应的具有集合论极限的 Horn 逻辑程序序列,则此归纳逻辑程序是收敛的,并且如果该 Horn 逻辑程序序列关于例子集合序列的极限是极限正确的,则此归纳逻辑程序是极限正确的,还说明 GOLEM 系统不是极限正确的. 为了解决这个问题,提出了一个极限正确的称为优先 GOLEM 系统的归纳逻辑系统,并证明了在一定的限制下,优先 GOLEM 系统的算法是极限正确的.

关键词:归纳逻辑程序;机器学习;极限行为
中图分类号:TN911.7 **文献标识码**:A **文章编号**:1673-4785(2007)04-0009-05

Limit behavior of prioritized inductive logic programs

MA Shi-long[1,2] ,SUI Yue-fei[2] ,XU Ke[1]
(1. School of Computer Science,Beihang University,Beijing 100083,China; 2. Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100080,China)

Abstract:Limit behavior of inductive logic programs is an important research topic, but it has not been deeply explored until now. When running incremental or online inductive learning algorithms on a computer, limit behavior should be taken into account. An example is given to show that some inductive learning algorithms may produce errors in the long run if limit behavior is not considered. An inductive logic program is convergent if, given an increasing sequence of example sets, the inductive logic program creates a corresponding sequence of Horn logic programs which have the set-theoretic limit. Furthermore, it is limit-correct if the limit of the sequence of the Horn logic programs is correct with respect to the limit of the sequence of the example sets. Two examples show that the GOLEM system is not limit-correct. Finally, we propose an inductive logic program which is limit-correct, called the prioritized GOLEM system. It is proved that the prioritized GOLEM is limit-correct under certain constraints.

Keywords:inductive logic program; machine learning; limit behavior

随着信息量呈指数倍的增长,从大量信息中发现有用的知识变得越来越重要. 归纳逻辑程序是一种从例子中学习一般理论的程序. 在增量式的学习中,例子总是一个接一个给出的. 每获得一个新的例子,从以前的例子中学习到的理论需要不断地更新以满足目前出现的例子. 这样就能得到一个理论序列:

$$\prod_1, \prod_2, \cdots, \prod_n, \cdots.$$

有时会出现无穷多个例子,该过程无法停止,即不存在自然数 k 使得 $\prod_k = \prod_{k-1} = \cdots$,例如,如果把这些理论限制到 Horn 逻辑程序上,那么存在一个 Herbrand 解释 I 使人永远无法找到一个有限的程序 $\prod$,它的最小 Herbrand 模型等于 I,因为有限的 Horn 逻辑程序集合只满足可数多个条件时,该 Herbrand 解释集合是不可数的. 因此,还应该考虑理论的极限,使得对所有的例子理论都正确[1-3].

李未[4-5]独立地提出了将一阶理论的集合论极限引入到逻辑和计算机科学,并在收敛无穷计算中使用理论版本作为某些形式理论的逼近. 李未[4-5]定义了一阶理论的集合论极限,并随后给出了它的

收稿日期:2006-11-12.

归纳逻辑的形式系统. 精确地,给定一阶理论的序列 Π_n,$\{\Pi_n\}$的集合论极限,记为 $\Pi = \lim_{n\to\infty}\Pi_n$ 是一些语句的集合,满足:每条包含在 Π 中的语句几乎包含在每个 Π_n 中,且在无限多的 Π_n 中的每条语句也包含在 Π 中. 只有部分一阶理论序列存在集合论极限. 本文用极限表示集合论极限.

归纳逻辑程序的极限行为至今还是个未知的研究领域. 目前大部分软件和算法都是增量式的或在线的,需要长期运行. 当考虑此类软件和算法的正确性问题时,它们的极限行为也应该考虑. 本文研究的重点是归纳逻辑程序中的增量式归纳学习算法. 假设存在一个例子序列,并令 E_n 为 n 时刻的例子. 归纳学习算法 A 产生理论 $A(E_n)$,其中对每个 n,理论 $A(E_n)$ 关于 E_n 都是正确的. 然后将给出一个实例表明如果不考虑极限行为,有的归纳学习算法就不正确. 就极限行为而言,合理的归纳学习算法 A 应该满足下面的条件:

1) 收敛性:对给定的例子集合序列$\{E_n\}$使得 $E_1 \subseteq E_2 \subseteq E_3 \subseteq \cdots$,$\{A(E_n)\}$有集合论极限;

2) 极限正确性:$A(E_n)$的极限应该关于 E_n 的极限是正确的,即,对每个 $e \in \lim_{n\to\infty} E_n$ 有 $\lim_{n\to\infty} A(E_n) \vdash e$;

可以根据上面的条件考虑下面 ILP 系统,如 FOIL、GOLEM 和 MOBAL. 因为 FOIL 和 MOBAL 是无函数的,所以这里考虑的重点是 GOLEM 系统. 下面将讨论基于一个只含有有限多谓词符号的固定的逻辑语言.

考虑了 Horn 逻辑程序的极限行为,并对以下定理进行了证明.

定理 1 给定 Horn 逻辑程序序列$\{\Pi_n\}$,若 $\Pi = \lim_{n\to\infty}\Pi_n$ 存在且对每个足够大的 n,Π_n 满足下列假设[7]:

对所有 Π_n 的子句,每个出现在体中的项都会出现在头中,则

$$\lim_{n\to\infty} M(\Pi_n) = M(\lim_{n\to\infty}\Pi_n),$$

式中:M 是一个算子使得对任意的 Horn 逻辑程序 Π,$M(\Pi)$是 Π 的最小 Herbrand 模型.

为了说明归纳逻辑程序的极限正确性,假设归纳逻辑程序满足收敛性. 给定一个归纳逻辑程序 A,若对所有的正例集合 E_n,$A(E_n)$为关于 E_n 正确的 Horn 逻辑程序,即,$M(A(E_n)) \supseteq E_n$,且 $A(E_n)$满足假设,则根据定理 1,得

$$M(\lim_{n\to\infty} A(E_n)) = \lim_{n\to\infty} M(A(E_n)) \supseteq \lim_{n\to\infty} E_n.$$

所以 A 是极限正确的. 因此,为了使 A 满足极限正确性,还要设计 A 使得对任意的输入 E,$A(E)$就是满足假设的 Horn 逻辑程序. 下面举 2 个实例来说明目前的 GOLEM 系统产生的逻辑程序不满足假设. 将 GOLEM 系统修改为一个优先 GOLEM 系统,假设其中的文字上定义一个优先序. 详细地,令 G 和 P 分别为 GOLEM 算法和优先 GOLEM 算法,对任意的例子集合 E,用 $G(E)$ 和 $P(E)$ 分别表示由 GOLEM 系统和优先 GOLEM 系统产生的 Horn 逻辑程序. 则对例子集合 E,$G(E)$ 可能不会满足假设,但 $P(E)$ 能满足假设. 这样,P 就是极限正确的.

1 GOLEM 系统

类似于 Nienhuys-Cheng[9] 的项和公式上距离的定义,下面给出一个距离的不同定义.

定义 1 令 f 和 g 分别为 n 元和 m 元函数符号,距离 ρ 定义如下:

1) 对任意项 t,有 $\rho(t,t) = 0$;

2) 如果 $f \neq g$,则

$$\rho(f(t_1,\cdots,t_n), g(s_1,\cdots,s_m)) = 1.$$

3) $\rho(f(t_1,\cdots,t_n), f(s_1,\cdots,s_n)) =$

$$\frac{\max\{\rho(t_i,s_i) \mid 1 \leq i \leq n\}}{\max\{\rho(t_i,s_i) \mid 1 \leq i \leq n\} + 1},$$

式中:$t_1,\cdots,t_n,s_1,\cdots,s_m$ 为项.

上面定义的距离与 Nienhuys-Cheng 定义的不同之处在于距离的值可以是一个简单的形式 $1/m$,其中 m 为某自然数. 这样的距离用于图论中树之间的距离. 每个项都可以看成是一棵树 T_t. 例如,$t = f(t_1,\cdots,t_n)$,树 T_t 具有标有符号 f 的根,及 n 个子树 $T_{t_1},\cdots,T_{t_n}$. 如果 T_t 和 $T_{t'}$ 在深度 m 上相同,则 t 和 t' 在深度 m 相同.

设有 2 个子句 C_1 和 C_2,为了计算 C_1 和 C_2 的最小一般概化(the least general generalization of C_1 and C_2),用 $\mathrm{lgg}(C_1,C_2)$表示,给出如下计算过程:

步骤 1 设项 $t = f(t_1,\cdots,t_n)$和项 $s = g(s_1,\cdots,s_m)$,

$$\mathrm{lgg}(t,s) = \begin{cases} v, & f_1 \neq f_2, \\ f_1(\mathrm{lgg}(t_{11},t_{21}),\cdots,\mathrm{lgg}(t_{1n},t_{2n})), & f_1 = f_2. \end{cases}$$

式中:v 为任意新的变量.

步骤 2 给定 2 个文字 $l_1 = (\neg)^{i_1} p_1(t_{11},\cdots,t_{1n})$和 $l_2 = (\neg)^{i_1} p_2(t_{21},\cdots,t_{2m})$,

$$\mathrm{lgg}(l_1,l_2) = \begin{cases} \text{无定义}, & i_1 \neq i_2 \text{ 或 } p_1 \neq p_2, \\ (\neg)^{i_1} p_1(\mathrm{lgg}(t_{11},\cdots,t_{21}),\cdots,\mathrm{lgg}(t_{1n},t_{2n})), & i_1 = i_2 \text{ 且 } p_1 = p_2. \end{cases}$$

式中:$i_1 = 0$ or 1, $(\neg)^0 = $, $(\neg)^1 = \neg$.

步骤 3　给定 2 个子句 $C_1 = \{l_{11}, \cdots, l_{1n}\}$ and $C_2 = \{l_{21}, \cdots, l_{2m}\}$,

$\mathrm{lgg}(C_1, C_2) =$

$\{\mathrm{lgg}(l_{1i}, l_{2j}) : 1 \le i \le n, 1 \le j \le m, \mathrm{lgg}(l_{1i}, l_{2j})$ 有定义$\}$.

设有 2 个子句集合 A, B,定义 $\mathrm{lgg}(A, B) = \{\mathrm{lgg}(C_1, C_2) : C_1 \in A, C_2 \in B, \mathrm{lgg}(C_1, C_2)$ 存在,

$$d(C_1, C_2) = \min\{d(C_1, C_2) : C_2 \in B\}\}.$$

Muggleton 和 Feng[8]证明:如果 Γ 是基文字的有限集合,那么关于 $\prod$ 的子句 C_1 和 C_2 的 rlgg 为 $\prod \to C_1$ 和 $\prod \to C_2$ 的 lgg. 基于这个性质,给出了 ILP 学习系统 GOLEM,该系统是唯一建立在相关的最小一般概化概念基础上的学习系统.

GOLEM 系统:

假设有逻辑系统 Γ(即背景知识)和 2 个实例(2 个基原子) E_1 和 E_2 使得 $\Gamma \nvdash E_1$ 且 $\Gamma \nvdash E_2$. 构造一个与 Γ 有关的 E_1 和 E_2 的lgg C,有 $\Gamma \wedge C$、$E_1 \wedge E_2$,且 C 在 E_1 和 E_2 的推导中只用到一次. 令 $\Gamma = \{p_1, \cdots, p_n, \cdots\}$.

定义 $C_1 = ((\neg p_2 \vee \neg p_1 \vee \cdots) \vee E_1)$, $C_2 = ((\neg p_1 \vee \neg p_2 \vee \cdots) \vee E_2)$ 和集合 $C = \mathrm{lgg}(C_1, C_2)$.

定义 2　给定公式集合的序列$\{A_n\}$,若满足

$$\overline{\lim_{n\to\infty}} A_n = \underline{\lim_{n\to\infty}} A_n,$$

其中 $\overline{\lim\limits_{n\to\infty}} A_n = \{P : \exists^{\infty} n (p \in A_n)\}$,

$$\underline{\lim_{n\to\infty}} A_n = \{p : \exists n_0 \forall n \ge n_0 (p \in A_n)\},$$

式中:$\exists^{\infty} n$ 表示存在无限多的 n 则$\{A_n\}$的集合论极限存在,记为 $\lim\limits_{n\to\infty} A_n$.

2　GOLEM 系统的极限正确性

本节考虑 GOLEM 系统的极限正确性,将给出 2 个实例说明 GOLEM 不是极限正确的,且对实例的顺序是敏感的.

令 p 为谓词, $p(x)$ 表示 x 为一个偶数; s 为后继函数,即 $s(x)$ 为 x 的后继. 令

$$E_n = \{p(0), p(s^2(0)), \cdots, p(s_{2n}(0))\}.$$

有对 E_n 产生不同理论的 2 种归纳学习算法.

情况 1　给定 E_n,其中一个归纳学习算法产生 $T_n = \{p(0); p(x) \to p(s^2(x))\}$,则 T_n 为 Horn 逻辑程序, T_n 的最小 Herbrand 模型为

$$M_n = \{p(0), p(s^2(0)), \cdots, p(s^{2m}(0)), \cdots\}.$$

可以得出:

$$T = \lim_{n\to\infty} T_n = T_1;$$

$$M = \lim_{n\to\infty} M_n = M_1.$$

T 的最小 Herbrand 模型为 M.

情况 2　给定 E_n,另一个归纳学习算法产生 $S_n = \{p(s^{2n}(0)); p(s^n(x)) \to p(x)\}$. 则 S_n 为一个 Horn 逻辑程序且 S_n 的最小 Herbrand 模型为 E_n,记做 N_n,但因为

$$S = \lim_{n\to\infty} S_n = \{p(s^2(x)) \to p(x)\};$$

$$N = \lim_{n\to\infty} N_n = \lim_{n\to\infty} E_n = M_1.$$

S 的最小 Herbrand 模型,称为 $M(S)$,等于空集. 则有

$$M(S) \ne N.$$

因此对任意 $e \in N$,有 $S \nvdash e$.

例 1　假设 E_n 如上所述,则 GOLEM 系统产生如下的 Horn 逻辑程序序列:

$T_0 = \{p(0)\}$;

$T_1 = \{p(0); \pi_1\}$;

$T_2 = \{p(0); \pi_1; \mathrm{lgg}(\pi_1, \pi_2)\} =$

$\{p(0); \pi_1, p(x) \to p(s^2(x))\} =$

$\{p(0); p(x) \to p(s^2(x))\}$,

$T_3 = T_2$,

…

$T_n = T_2$,

…

式中:

$$\pi_2 = \{\neg(\neg p(0) \vee p(s^2(0))), p(s^4(0))\} = \{p(0), p(s^4(0))\} \wedge \{\neg p(s^2(0)), p(s^4(0))\},$$

$$\pi_1 = \{\neg p(0), p(s^2(0))\},$$

$$\mathrm{lgg}(\pi_1, \pi_2) = \{p(s^2(x)), \neg p(x)\}.$$

修改 $p(s^{2n}(s)$ 出现的顺序,然后看看 GOLEM 系统得到的结果.

例 2　假设

$$E_0' = \{p(s^4(0))\},$$

$$E_1' = E_0 \cup \{p(s^2(0))\},$$

$$E_2' = E_1 \cup \{p(0)\},$$

$$\cdots$$

$$E_{3k}' = E_{3k-1} \cup \{p(s^{6k+4}(0))\},$$

$$E_{3k+1}' = E_{3k} \cup \{p(s^{6k+2}(0))\},$$

$$E_{3k+2}' = E_{3k+1} \cup \{p(s^{6k}(0))\}, \cdots$$

则 GOLEM 系统会产生如下的 Horn 逻辑程序序列:

$$S_0 = \{p(s^4(0))\};$$

$$S_1 = \{p(s^4(0)); \gamma_1\};$$

$$S_2 = \{p(s^4(0)); \gamma_1; \mathrm{lgg}(\gamma_1, \gamma_2)\} =$$

$$\{p(s^4(0)); \gamma_1; p(s^2(x)) \to p(x)\} =$$

$$\{p(s^4(0)); p(s^2(x)) \to p(x)\},$$

$$\cdots$$

$$S_{3k+i}=\{p(s^{6k+4-2i}(0));p(s^2(x))\rightarrow p(x)\},$$

$$\cdots$$

式中:

$$\gamma_1=\{\neg p(s^4(0)),s^2(0))\},$$

$$\gamma_2=\{p(s^4(0)),p(0)\}\wedge\{\neg p(s^2(0)),p(0)\},$$

$\mathrm{lgg}(\gamma_1,\gamma_2)=\{p(x),\neg p(s^2(x))\}$.

从上述可知,该实例说明 GOLEM 系统不是极限正确的,且对例子的顺序是敏感的.

3　优先 GOLEM 系统

笔者[7]提出了关于 Horn 逻辑程序的假设,用以保证 Horn 逻辑程序序列的极限的最小 Herbrand 模型就是逻辑程序的最小 Herbrand 模型的极限,且前一个极限的最小 Herbrand 模型关于 Horn 逻辑程序序列是不变的.

定义 2　一个子句中,若每个出现在子句体中的子项都出现在该子句头中,则称该子句是简单的.一个逻辑程序,若其中的每个子句都是简单的,则称该逻辑程序是简单的.

前文的假设要求 Horn 逻辑程序是简单的.给定例子集合 E,将 GOLEM 系统修改为新的算法 P 使得 $P(E)$ 是简单的.

先定义文字上的优先关系 $\prec$.假设有 2 个文字 l 和 l',如果 l 中出现的每个子项都在 l 中出现,则说 l 比 l 优先级高,记做 $l\prec l'$.

命题 1　$\prec$ 是一个前序,即 $\prec$ 是自反的和传递的.

优先 GOLEM 系统:

假设给定一个逻辑程序 Γ(即一个背景知识)和 2 个例子(2 个基原子) E_1 和 E_2 使得 $\Gamma\nvdash E_1$ 和 $\Gamma\nvdash E_2$ 成立.构造一个与 Γ 有关的 E_1 和 E_2 的 lgg C,有 $\Gamma\wedge C\vdash E_1\wedge E_2$ 且 C 在 E_1 和 E_2 推导中用了一次令 $\Gamma=\{p_1,\cdots,p_n,\cdots\}$.

定义

$$C_1=((\neg q_1\vee\neg q_2\vee\cdots)\vee e_1),$$

$$C_2=((\neg q_1'\vee\neg q_2'\vee\cdots)\vee e_2),$$

集合 $C=\mathrm{lgg}(C_1,C_2)$,其中 $q_1,q_2,\cdots,e_1$ 为带序 $\prec$ 的集合 $\{p_1,p_2,\cdots,E_1\}$,即,对每个 i 有 $\{q_1,q_2,\cdots,e_1\}=\{p_1,p_2,\cdots,E_1\}$,且 $e_1\not\prec q_i$;类似地可定义 $\{q_1',q_2',\cdots,e_2\}$.

序列上的优先 GOLEM 系统:

给定例子集合序列 $\{E_n\}$,在步骤 n 输入 E_n.若对任意 $i<n$,有 $E_n\not\prec E_i$,则可用 GOLEM 系统直接产生一个 Horn 逻辑程序,记做 $A(E_n)$;否则,可以找到最小的 $i<n$ 使得 $E_n\prec E_{i+1}$,然后可以用背景知识 $P(E_i)$ 把 GOLEM 系统应用到 $\bigcup_{j>i}^{n}E_j$ 上.令 $P(E_n)$ 为结果.

定理 2　给定例子集合 E, $P(E)$ 是简单的.

证明　根据前序的定义,当一个子句 π 列举在 $P(E)$ 中时,π 的头在 E 中总是有最低优先级.这就保证了 π 是简单的.

这样,可以得出如下定理.

定理 3　P 是收敛且极限正确的.即给定一个正例集合的增长序列 $\{E_n\}$,1) $\{P(E_n)\}$ 是有集合论极限的 Horn 逻辑程序序列;2) $\lim\limits_{n\to\infty}P(E_n)$ 是关于 $\lim\limits_{n\to\infty}E_n$ 极限正确的.

证明　根据前文的讨论只需要证明 1).根据前序 $\prec$ 的定义,可知 $\prec$ 是良定的.给定一个实例 e,只存在有限多个 e 使得 $e'\prec e$.假设对某 n, e 是在 E_n 中可列举的,则子句 π 产生.对 $n'>n$,只有当在 E_n 中实例 $e'\prec e$ 可列举时,π 才能从 $P(E_n)$ 中提取.这样对无限多 n, π 不能在 $P(E_n)$ 中列举且从 $P(E_n)$ 中提取出.因此 $\{P(E_n)\}$ 具有集合论极限.

4　结束语

当输入例子集合 E 时,优先的 GOLEM 系统产生一个简单的 Horn 逻辑程序 $P(E)$.简单 Horn 逻辑程序有很多好的普通的 Horn 逻辑程序没有的性质.优先的 GOLEM 系统是基于例子的语法性质,即,文字上的优先序 $\prec$,该序可以使优先 GOLEM 系统在多种应用中变得有用.

将来的工作可以基于一个逻辑语言的项或公式上定义的距离.这种距离定义可以采用 Fitting 或 Nienhuys-Cheng 给出的距离.然后可以定义项或公式的 Cauchy 序列.然后根据例子集合的 Cauchy 序列和 Horn 逻辑程序定义收敛性和极限正确性.推测优先 GOLEM 系统还满足定义在 Cauchy 序列上的收敛性和极限正确性.

参考文献:

[1] BERGADANO F, GUNETTI D. Inductive logic programming: from machine learning to software engineering [M]. London: The MIT Press, 1996.

[2] DAHR M. Deductive eatabases: theory and applications [M]. Boston: International Thomson Computer Press, 1997.

[3] FITTING M. MeETRIC methods, three examples and a theorem[J]. J of Logic Programming, 1994, 21: 113 - 127.

[4] LI W. An open logic system[J]. Science in China (Scien-

tia Sinica) (series A) ,1992(10) :1103 - 1113.

[5]LI W. A logical framework for inductive inference and its rationality [A]. Advanced Topics in Artificial Intelligence ,LNAI 1747[C]. Berlin : Springer ,1999.

[6]LLOYD J W. Foundations of logic programming [M]. Berlin : Springer-Verlag ,1987.

[7]MA S, SUI Y, XU K. The limits of the horn logic programs[A]. Lecture Notes in Computer Science (LNCS) [C]. [S.l.] ,2005.

[8]MUGGLETON S, FENG C. Efficient inductive of logic programs [A]. Proc of the First Conf on Algorithmic Learning Theory[C]. Tokyo ,1990.

[9]NIENHUYS-CHENG S H. Distances and limits on herbrand interpretations [A]. Proc of the 8th International Workshop on Inductive Programming ,LNAI 1446 [C]. Springer ,1998.

[10]PLOTKIN G. A note on inductive generalization [J]. Machine Intelligence , 1970 ,5 :153 - 163.

作者简介:

马世龙,男,1953 年生,教授,主要研究方向为网络环境下计算研究、海量信息处理计算模型研究、网格计算技术及其应用研究,发表学术论文 30 余篇.

E-mail :slma @nlsde. buaa. edu. cn.