

适应噪声环境的解释学习算法*

张旗 石纯一

(清华大学计算机系 北京 100084)

摘要 在现实世界里, AI 系统难免受到噪声的影响. 系统有效工作与否取决于它对噪声的敏感性如何. 解释学习 EBL (explanation-based learning) 也不例外. 本文探讨了在例子受到噪声影响的情况下, 解释学习的处理问题, 提出了一个算法 NR-EBL (noise-resistant EBL). 与现有的解释学习方法不同, NR-EBL 在训练例子含有噪声时仍然可以学习, 以掌握实际的问题分布; 和类似的工作不同, NR-EBL 指出了正确识别概念对于噪声规律的依赖性, 试图从训练例子集合发现和掌握噪声的规律. 可以相信, 在识别概念时, 借助于对噪声规律的认识, NR-EBL 可比 EBL 和类似工作有更高的识别率. NR-EBL 是解释学习和统计模式识别思想的结合. 它把现有的解释学习模型推广到例子含有噪声的情形, 原来的 EBL 算法只是它的特例.

关键词 解释学习, 噪声, 模式识别.

在现实世界里, 数据中含有某种噪声是普遍现象, AI 系统难免受到噪声的影响. 而噪声又往往使设计精巧的理论系统在实际问题面前显得非常脆弱. 系统有效工作与否取决于它对噪声的敏感性如何, 因此, 考虑到噪声处理的那些 AI 系统能更好地胜任复杂恶劣的现实环境, 实现所设计的目标.

解释学习, 使用和问题有关的领域理论, 构造能够证明训练例子属于目标概念的成员的_{解释}, 再从其中得出一般性的宏观规则 (Macro Rule), 或称为组块规则 (Chunking Rule), 从而以后可借助所学的宏观规则迅速地识别同类的例子, 提高系统的效率.^[1,2]

解释学习需要领域理论和例子, 因此在现实世界中它就必然会遇到噪声的影响. 事实上, 解释学习对于噪声是很敏感的. 例子中的噪声会有 2 方面的影响. 在学习阶段, 一方面噪声改变了训练例子的面貌, 可能使其不再是目标概念的成员, 掩盖了实际的问题分布 (Problem Distribution) 即例子分布; 另一方面, EBL (explanation-based learning) 在尝试完领域理论所有可能的证明之后才发现这点, 结果既没有学到相应于实际问题分布的宏观规则, 又开销很多时间. 在使用阶段, EBL 只是用宏观规则去判断例子, 对于受到噪声影响、类别改变的例子, EBL 给出的判断是错误的, 不能做出一定的矫正. 所以, 有必要扩展现有的解释学习模型, 使它对于受到噪声影响的例子, 在学习阶段仍然可以进行学习, 在使用中又有较高的

* 作者张旗, 1963 年生, 讲师, 主要研究领域为人工智能, 机器学习, 专家系统. 石纯一, 1935 年生, 教授, 博士生导师, 主要研究领域为人工智能应用基础, 知识工程.

本文通讯联系人: 张旗, 北京 100084, 清华大学计算机系

本文 1995-04-20 收到修改稿

识别正确率.这也是本文期望达到的2个目标.

基于这样的思想,本文提出了一个学习算法 NR-EBL(noise-resistant EBL),它扩展了现有的解释学习模型,考虑了对噪声的处理,可以更好地适应现实环境.

1 一些考虑

领域理论完善与否制约着 EBL 的使用.改善不完善的领域理论(Imperfect Domain Theory)是解释学习研究的一个重要方面.^[1,3]受到噪声影响的领域理论是有缺陷的,表现在领域理论不正确、不一致等.本文只考虑完善的领域理论,仅例子有噪声.这么做基于2个理由.首先,受到噪声影响的领域理论仍然属于不完善领域理论的研究,这方面已有很多工作.其次,在例子中含有噪声的情况下,倘若再考虑不完善的领域理论,则会使研究变得复杂,因为对于 EBL 的分类错误很难指出是例子还是领域理论造成的,同时还要考虑改善有缺陷的领域理论.关于改善不完善的领域理论,可见文献[4~8]等.

本文强调训练例子的集合要能反映实际的问题分布和噪声规律.解释学习隐含的一个思想是训练例子集合能够代表以后出现的例子的分布状况,从而所学的宏规则就比较接近实际问题,被最大程度地使用,提高系统的整体效率.从这个意义上说,EBL 是总结经验、产生技能知识的学习;经验与现实世界的远近直接影响系统对将来问题的处理.同样,在例子中有噪声时,我们还要求训练例子集合能够反映出噪声的规律,让 EBL 设法找出这些规律.这时,EBL 所获得的经验不仅包括学到的宏规则,而且还包括对噪声规律的认识.难以想象,不了解噪声的规律,仅靠宏规则会有可靠的识别.本文强调噪声的规律性,给出的 NR-EBL 算法适用于噪声有规律的情形.

本文假设例子中只有属性噪声(Attribute Noise),没有例子分类噪声(Positive/Negative Noise),各个属性上的噪声是相互独立的.

2 NR-EBL 算法

NR-EBL 在学习和概念识别上都对现有的 EBL 模型进行了扩展.

2.1 学习

学习阶段的目标是不仅发现和学到与实际问题分布,得到一些宏规则,而且还要发现和总结出有关噪声的规律,得到相应的概率模型.

目标概念的训练例子可以描述为: $class(a_1, a_2, \dots, a_m)$

其中, $class$ 表示教师给出的例子分类,取值“+”或“-”(分别表示正例和反例). a_1, a_2, \dots, a_m 是例子分别在属性 A_1, A_2, \dots, A_m 上的取值.记训练例子集合是 S .

S 中的例子可分为4类:(1)没有受到噪声影响的正例,学习时系统仍能由领域理论得出例子是正例的解释;(2)没有受到噪声影响的反例,学习时系统仍能由领域理论得出例子是反例的解释;(3)受到噪声影响的正例,学习时系统由领域理论得出例子是反例的解释;(4)受到噪声影响的反例,学习时系统由领域理论得出例子是正例的解释.

设立一些变量进行计数,以得出噪声因素的有关规律.设 n 表示 S 中例子的总数, $n_c, n_{\bar{c}}, n_{c\bar{c}}, n_{\bar{c}c}$ 分别表示上述4种例子的个数, n_+ 表示 S 中可由领域理论判为正例的例子个数, n_-

表示属性 A_i 受到噪声影响的次数 ($1 \leq i \leq m$).

算法 NR-EBL 在学习阶段确定含有噪声的属性并记录噪声出现的频率,从受到噪声影响的正例中学到相应的宏规则.

算法 NR-EBL(Learning):

- (1) 把所有的计数变量置为零;
- (2) 对 S 中的每一个例子,做如下处理:

1) $n \leftarrow n + 1$. 用领域理论去解释例子. 如果系统给出的分类与例子的 $class$ 一致,转到 2), 否则转 3).

2) 如果 $class$ 为 +, 则由例子学到一条宏规则, $n_c \leftarrow n_c + 1, n_{cc} \leftarrow n_{cc} + 1$; 否则, 采用归纳学习的方法对反例进行学习, 得到有关反例的规则, $n_{\bar{c}} \leftarrow n_{\bar{c}} + 1$. 转(2).

3) 如果 $class$ 为 +, 则该例的某些属性含有噪声. 对于与领域理论中解释树的叶子不吻合的那些属性 $A_i, n_i \leftarrow n_i + 1$, 然后由这棵解释树学到例子未受噪声影响时所对应的宏规则, $n_{\bar{c}} \leftarrow n_{\bar{c}} + 1$; 否则, $n_{cc} \leftarrow n_{cc} + 1, n_{\bar{c}\bar{c}} \leftarrow n_{\bar{c}\bar{c}} + 1$, 再对于每个属性 $A_i, n_i \leftarrow n_i + k$.

(3) 计算噪声的概率模型:

$$p_i \cong \frac{n_i}{n} \quad (1 \leq i \leq m),$$

$$p_c \cong \frac{n_c}{n}, p_{cc} \cong \frac{n_{cc}}{n}, p_{\bar{c}} \cong \frac{n_{\bar{c}}}{n}, p_{\bar{c}\bar{c}} \cong \frac{n_{\bar{c}\bar{c}}}{n}, p_{c\bar{c}} \cong \frac{n_{c\bar{c}}}{n},$$

$$p(C|C) \cong \frac{p_{cc}}{p_c}, p(\bar{C}|\bar{C}) \cong \frac{p_{\bar{c}\bar{c}}}{1-p_c}, p(C|\bar{C}) \cong \frac{p_{c\bar{c}}}{1-p_c}, p(\bar{C}|C) \cong \frac{p_{\bar{c}c}}{p_c}$$

由于没有关于目标概念的反例的领域理论,所以在(2)的 3)里当 $class$ 为一、系统判断例子为正例时,无法推测出哪些属性有噪声,故对每个 n_i 引入可调参数 $k, 0 \leq k \leq 1$. 因为训练例子集合能反映噪声的规律,因而可用有关的频率去近似相应的概率. 概率 $p_i, p_c, p_{cc}, p_{\bar{c}}, p_{\bar{c}\bar{c}}, p_{c\bar{c}}$ 的意义不难看出. 后验概率 $p(C|C), p(\bar{C}|\bar{C}), p(C|\bar{C}), p(\bar{C}|C)$ 分别表示了由领域理论得出受到噪声影响的例子分布之后,人们对于其原来类别的可能性估计.

2.2 概念识别

在识别阶段, NR-EBL 算法先借助现有的 EBL 算法对例子的类别做出初步的估计,但更重要的是利用学习阶段得到的概率模型和例子的特点确定是否对初步的结果进行修正.

首先计算出 $p_{\max} = \max\{p_i \mid 1 \leq i \leq n\}$.

算法 NR-EBL(concept recognition):

(1) 置集合 D 为空. 对于所给的例子描述, 系统去识别. 若成功, 转(2); 否则转(3).

(2) 用在学习阶段得到的反例的规则去解释例子, 对于例子中不能与规则吻合的那些属性 $A_i, D = D \cup \{i\}$. 计算:

$$B_1 = 1 - \frac{p(\bar{C}|C)}{p(C|C)} \times f,$$

其中 f 取值: 如果 $\|D\| = m$, 则 $f = 0$; 如果 $\|D\| \neq m$, 则 $f = \prod_{i \in D} p_i / (p_{\max} - g_1)$ 如果 $B_1 \geq \alpha$, 系统输出 C ; 否则输出 \bar{C} .

(3)对于例子中不能与领域理论的规则吻合的那些属性 $A_i, D = D \cup \{i\}$. 计算:

$$B_2 = 1 - \frac{p(C|\bar{C})}{p(\bar{C}|\bar{C})} \times f,$$

其中 f 取值: 如果 $\|D\| = m$, 则 $f = 0$; 如果 $\|D\| \neq m$, 则 $f = \prod_{i \in D} p_i / (p_{\max} - g_2)$. 若 $B_2 \geq \beta$, 系统输出 \bar{C} ; 否则输出 C .

算法中的 g_1 和 g_2 及 α 和 β 都是可调参数.

2.3 讨 论

上面设计的识别例子的公式, 目的在于当宏规则和领域理论对例子做出判断之后, 实现对判断的结果进行修正. 由于关于噪声规模较小的假设, 有 $p(C|C) > p(\bar{C}|C)$ 和 $p(\bar{C}|\bar{C}) > p(C|\bar{C})$. 要注意到, 上述公式判断例子的方法和统计模式识别的作法很类似. 如果取 $f = 1, \alpha$ 和 β 都为零, 则完全是统计模式识别. 倘若这样, 就没什么修正的问题了, 宏规则和领域理论判出什么就是什么. 但从统计的角度讲, 这么做分别有 $p(\bar{C}|C) \times 100\%$ 和 $p(C|\bar{C}) \times 100\%$ 的错误, 而我们的目的就是要降低这些错误率.

对于是否修改结果, 必须考虑这些因素: (1) 如果 $p(\bar{C}|C)/p(C|C)$ 或 $p(C|\bar{C})/p(\bar{C}|\bar{C})$ 很小, 则修改的可能性就小. (2) 如果有 $\|D\| = m$, 说明相差极大, 不应修改. (3) 若 D 中某个属性 A_i 对应的概率 p_i 为零, 则说明相反结果成立的必要条件之一不可能发生, 因此不必修改. (4) $\|D\|$ 越大, 则修改的可能性越小, 因为多个独立的概率事件同时发生的可能性比较小; 反之, 则增大修改的可能性. (5) 若 $\|D\|$ 很小, 且 D 的元素对应的概率 p_i 在所有属性对应的概率之中是最大或很大的, 则可能对结果进行修改.

基于这些考虑, 设计了上述概念识别的公式. 取 g_1 和 g_2 比零略大一点 ($0 < g_1 < 1, 0 < g_2 < 1$), 取 α 和 β 比零大一点 ($0 < \alpha < 1, 0 < \beta < 1$). 参数的具体取值依赖于实际问题. 公式是否合适、是否需要重新设计还要看实验的结果.

2.4 与相关工作的比较

关于 EBL 在例子数据有噪声情况下的研究, 可见的只有 Mani 的工作.^[9] Mani 给出了一个算法 NH. 本文关于噪声的假设均与文献[9]相同. NH 没有涉及从有噪声的例子如何学习的问题, 也没有考虑从训练例子集合获取噪声规律. NH 的主要思想是: 在识别概念时, NH 计算例子和宏规则的差异(非匹配的子句的个数); 若差异为零, 则判为正例; 如果非零且大于某个阈值, 则例子是反例; 若小于这个阈值, 此时如果领域理论是正确的或过分特殊的(Overly Specific), 则例子是正例, 否则做一些其它处理. 和 NH 相比, NR-EBL 有几个优点. 首先, NH 用子句的差异做判断的尺度比较牵强, 因为 2 个较近的概念的差异很可能比较小, 再有, 如果唯一一个有差异的属性上根本不会有噪声, 则也不能用差异作为判断的标准. 而 NR-EBL 结合噪声的规律进行概念识别就比较科学可信. 其次, 当差异为零时, NH 判为正例是要冒风险的, 因为可能是噪声把反例变成了正例. 结合以往的概率考虑就比较可靠, 这也正是统计模式识别的作法. 第 3, NH 把领域理论分为过分一般(Overly General)和过分特殊(或正确)两类, 这么做不全面, 因为领域理论完全可能 2 种情况都有, 或者还有其它不完善特点.^[3] 此时, 应该走 NH 的哪个分支? 再者, 在 NH 里还必须把领域理论的上述两类特点事先输入, 否则无法确定走算法的哪个分支. 所以较好的研究方法是假设领域理论是完善的. 当然, 在领域理论为不完善的情况下也能较好地处理噪声是比较理想

的,然而这非常困难.第4,在训练例子有噪声时,NR-EBL 仍然由领域理论做指导,学到相应的宏规则.NH 没有探讨这一问题.

可以把 NR-EBL 算法看成是统计模式识别和 EBL 的结合.EBL 在识别概念上有较高的效率;然而只靠 EBL,不了解和掌握噪声的规律,识别正确率不会高.NR-EBL 吸收了统计模式识别,从训练集中得出噪声的规律,以指导以后的识别.但 NR-EBL 并没有完全照搬统计模式识别的方法.如果仅用统计模式识别的办法,不利用现有的完善的领域理论和宏规则,则无法识别或正确率很低.

此外,要注意到,现有的 EBL 算法^[1,2]是 NR-EBL 在噪声为零时的特例.此时, $p(\bar{C}|C)$ 和 $p(C|\bar{C})$ 都是零,NR-EBL 的输出和现有的 EBL 算法完全一致.所以 NR-EBL 算法扩展了现有的 EBL 模型.

NR-EBL 方法的缺点是要求训练例子的集合比较大,以满足统计的要求,识别概念的效率较现有的 EBL 要低(但和 NH 差不多),这是为提高识别正确率所必须付出的代价.

3 小 结

本文探讨了在例子受到噪声影响的情况下 EBL 的处理方法问题,提出了一个算法 NR-EBL.与现有的解释学习方法不同,NR-EBL 在训练例子有噪声时仍然可以进行学习,较好地掌握实际的问题分布;和相关的工作^[9]不同,NR-EBL 指出了正确识别概念对于噪声规律的依赖性,试图从训练例子集合发现和掌握噪声的规律,借助于对噪声规律的认识来提高识别率,所以 NR-EBL 会比 EBL 和文献^[9]中提出的算法有更高的识别率.NR-EBL 是解释学习和统计模式识别的结合,把现有的解释学习模型扩展到例子含有噪声的情形,原来的 EBL 学习算法只是它在噪声为零时的特例.

EBL 学习模型仍然需要深入全面的研究^[10],作为进一步的工作,还可再探讨例子中含有属性噪声时的解释学习算法,探讨同时在领域理论具有某种不完善特征情况下例子噪声的处理问题.

参考文献

- 1 Mitchell T M, Keller T M, Kedar-Cabelli S T. Explanation-based learning: a unifying view. *Mach. Learn.*, 1986,1(1):47~80.
- 2 De Jong G F, Mooney R J. Explanation-based learning: an alternative view. *Mach. Learn.*, 1986,1(2):145~176.
- 3 Elman T. Explanation-based learning: a survey of programs and perspectives. *ACM Comput. Surv.*, 1989,21:163~221.
- 4 Rajamoney S A. Explanation-based theory revision: an approach to the problems of incomplete and incorrect theories. Dissertation, Urbana, Ill. Univ. of Illinois, 1988.
- 5 Tadepalli P. Lazy explanation-based learning: a solution to the intractable theory problem. *IJCAI-89*, 1989, 694~700.
- 6 Pazzani M J. Integrated learning with incorrect and incomplete theories. *IMI-88*, 1988. 291~297.
- 7 Qurston D, Mooney R J. Changing the rules: a comprehensive approach to theory refinement. *AAAI-90*, 1990, 815~820.
- 8 Cohen W W. Abductive explanation-based learning: a solution to the multipleinconsistent explanation problem.

Mach. Learn., 1992, 8:167~219.

- 9 Mani G. Handling noise in EBL: an abductive approach. Journal of Experimental and Theoretical Artificial Intelligence, 1992, 1:59~73.
- 10 Etzioni O. A structural theory of explanation-based learning. Artificial Intelligence, 1993, 60:93~139.

EXPLANATION-BASED LEARNING ALGORITHM FOR NOISY DATA ENVIRONMENT

Zhang Qi Shi Chunyi

(Department of Computer Science Tsinghua University Beijing 100084)

Abstract In the real world, AI systems are constantly and adversely influenced by noisy data. This is also true of EBL (explanation-based learning). This paper discusses how to cope with noisy data in explanation-based learning and proposes a NR-EBL (noise-resistant explanation-based learning) algorithm. Unlike existing algorithms, NR-EBL can learn macro rules and find the problem distribution when there is noise in training examples. Also unlike similar work, NR-EBL reveals the dependency of classifying examples correctly upon the regularities of noise and attempts to detect noise regularities from a set of training examples. With the help of knowledge of noise regularities, NR-EBL can have a higher rate of correct recognition than traditional algorithms and previous work. NR-EBL is the combination of explanation-based learning and statistical pattern recognition. Traditional algorithms are only special cases of NR-EBL when there is no noise in training examples.

Key words Explanation-based learning, noisy data, pattern recognition.