混合型学习模型 HLM 中的增量学习算法*

陈兆乾 周志华 李红兵 谢俊元

(南京大学计算机科学与技术系 南京 210093)

(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘要 混合型学习模型 HLM 将概念获取算法 HMCAP 和神经网络算法 FTART 有机结合,能学习多概念和连续属性,其增量学习算法建立在二叉混合判定树结构和 FTART 网络的基础上,在给系统增加新的实例时,只需进行一遍增量学习调整原结构,不用重新生成判定树和神经网络,即可提高学习精度,速度快、效率高. 本文主要介绍该模型中的增量学习算法.

关键词 混合模型,增量学习,神经网络.

中图法分类号 TP18

近年来,神经网络与符号学习相结合的混合学习模型逐渐成为机器学习领域的研究热点,混合型学习方法的研究已取得了不少成果,主要有 EITHER[1]等系统;KBANN[2]一类的算法;Utgoff 提出的耦合较为紧密的判定树与感知器相结合的方法[3];一种耦合度较低,用符号进行"高级决策"而用神经网络进行"低级决策"的方法[4]等等. HMCAP[5]是我们提出的一种用于多概念获取的符号学习算法,它采用属性值对表示法,为神经网络提供了良好的接口,可以处理连续属性. FTART[6]是我们提出的一种新的自适应谐振神经网络算法,该算法只需一遍学习,在样本数较少时也可实现对样本空间的有效划分. 混合型学习模型 HLM 的核心算法融合了 HMCAP 和 FTART 算法,其增量学习算法可以在最初生成的混合型判定树的基础上,根据每一个新增加的实例,对原有的树结构进行适当调整,使得不用重新生成一棵新的判定树就可以提高学习精度,速度快、效率高. HLM 模型已在 Windows95 环境下用 Visual C++4.0 编程实现,并已应用于"台风路径预报系统"中,取得了很好的效果.

1 HLM 模型概述

1.1 HLM 模型的结构

HLM 模型将基于概率论的符号学习与神经网络学习相结合,能从隶属于某个概念集的实例集中归纳出以混合型判定树表示的概念描述. 它首先调用 HMCAP 算法,利用离散

* 本文研究得到国家自然科学基金资助. 作者陈兆乾,女,1940 年生,教授,主要研究领域为机器学习和专家系统. 周志华,1973 年生,硕士生,主要研究领域为机器学习,神经网络. 李红兵,1973 年生,博士生,主要研究领域为知识工程,机器学习. 谢俊元,1962 年生,副教授,主要研究领域为图象处理,专家系统与机器学习.

本文通讯联系人:陈兆乾,南京 210093,南京大学计算机科学与技术系

本文 1997-03-18 收到修改稿

属性对一批训练实例进行分类,然后调用FTART网络对无法用离散属性精确划分的实例进行处理,利用这些实例的连续属性作进一步的学习.对新增实例,则调用其增量学习算法进行学习.应用结果表明,这样的处理方式对混合型学习系统来说可以取得很好的效果.

1.2 HMCAP算法

HMCAP是基于属性值对表示方式的符号学习算法,这种表示方式虽然结构性较弱,但可以方便地将属性值对转换为神经网络学习所需的输入模式对,有利于两者紧密耦合.

为了得到"最小花费"的判定树,HMCAP算法在概念和离散属性值对的组合中选择〈Attr,AttrValue,ConceptIndex〉,求得的相对于概念ConceptIndex的最好属性值对〈Attr,AttrValue〉;或者是尽可能多地覆盖隶属于概念ConceptIndex的例子,同时排斥一切隶属于其它概念的例子的属性值对;或者是以最大划分概率覆盖隶属于概念ConceptIndex的例子,同时尽可能多地排斥隶属于其它概念的例子的属性值对.然后根据每步求得的〈Attr,AttrValue,ConceptIndex〉生成判定树结构,为了获得尽可能高的精度,通常令归纳精度$Pr_i(s)=100\%$,达不到要求的结点就将引入FTART网络进行进一步学习.

限于篇幅,这里不再对HMCAP算法作进一步的介绍,具体内容见文献[5].

1.3 FTART算法

FTART算法[6]综合了Field Theory[7]和ART[8]以及ARTMAP[9]等算法的优点,其网络收敛速度极快,而且不存在陷入局部极小的问题,非常适合于在混合型学习中使用.FTART网络由4层神经元组成,其结构实际上是一个竞争型的3层网络分类器,可以形成对样本集的任意形状划分.使用这种结构,既具有神经网络分类器的优点,又可应用于广泛的问题领域中,有较强的适应性.

更为重要的是,与传统前馈型的BP算法不同,在给训练好的FTART网络增加新的输入模式时,不用再重新生成已有的网络结构,而只需在网络的第2层适当增加神经元,将它与部分已有神经元相连,并且适当地调整新增连接权,即可覆盖新增模式.FTART算法的这种增量学习能力为我们将要介绍的增量学习算法提供了强有力的支持.FTART算法的具体内容见文献[6].

2 增量学习算法

2.1 设计思想

为进行增量学习,我们要将新增实例与判定树进行匹配,但由于生成的混合型判定树中包含神经网络结点,因此在这里不能使用ID4[10],ID5R[11]中那种寻找最佳状态然后"上拉"的简单增量学习策略,否则将由于神经网络结点状态无法"上拉"而导致算法失败.因为HMCAP使用二叉混合型判定树,所以新增实例与判定树匹配的过程有其不同于ID4,ID5R的特征,即匹配过程在匹配到叶结点之前决不会停止,而无论到达哪一个叶结点,都将出现下面几种情况之一:

(1) 到达非神经网络结点,且新增实例的分类与该叶结点的分类相同;

(2) 到达非神经网络结点,新增实例的分类与该叶结点的分类不同,尚有可用于继续划分的离散属性;

(3) 到达非神经网络结点,新增实例的分类与该叶结点的分类不同,没有可用于继续划

分的离散属性;

(4) 到达神经网络结点,新增实例形成的输入模式已被原网络覆盖;

(5) 到达神经网络结点,新增实例形成的输入模式没有被原网络覆盖.

针对第(1)、(4)两种情况,我们不需要再做任何工作;对第(2)种情况,需要调用 HMCAP 算法对该叶结点进行进一步划分;对第(3)种情况需要生成一个新的神经网络结点;而对最后一种情况,则可以凭借 FTART 网络的增量学习功能,自动调整原网络的拓扑结构.

2.2 数据结构

混合型判定树中结点的数据成员及含义如表 1 所示. 结点类型中的 leaf 表示非神经网络的叶结点,NN 表示 FTART 网络结点;非叶结点的 nAttr 和 nAttrValue 记录由搜索最佳状态算法 BS[5]返回的用于划分的属性值对;叶结点的 nConceptIndex 中记录当前分类;属性表 AttrList 中用布尔值记录了可以用来对该结点进行划分的离散属性,如果 AttrList 表全为 TRUE 值,则表示不再有离散属性可用,该结点将调用 FTART 网络进行学习.

表 1　判定树中结点的数据成员及含义

结点类型	nNodeType {leaf,nonleaf,NN}
是否是左子树	nLtree {left,nonleft}
左子树指针	pLpointer
右子树指针	pRpointer
属性表	AttrList
BS 返回属性	nAttr
BS 返回属性值	nAttrValue
BS 返回概念号	nConceptIndex

已生成的混合型判定树的根结点由指针 pTree 确定. 新增实例 newExam 同样也组织成属性值对($\langle Attr_1, AttrValue_1\rangle, \langle Attr_2, AttrValue_2\rangle, \ldots, \langle Attr_n, AttrValue_n\rangle$)的形式,其类别为 nConcept.

2.3 增量学习算法描述

```
[初始化]
pCurrentNode=pTree;
[STEP1]
当前结点 pCurrentNode 是否是神经网络结点?
  是,则新增输入模式是否已被原网络覆盖?
       是,则返回;
       否则,调用 FTART 算法进行增量学习;
       返回;
  否则,GOTO STEP2;
[STEP2]
当前结点 pCurrentNode 是否是叶结点?
  是,则 newExam 的实际分类 nConcept=pCurrentNode→nConceptIndex?
       是,则返回;
       否则,GOTO STEP3;
  否则,GOTO STEP6;
[STEP3]
pCurrentNode 结点的 AttrList 表的内容是否全为 TRUE?
  是,则 pCurrentNode→nNodeType=NN;
    调用 FTART 算法生成一个新的神经网络结点;
```

返回;
否则,GOTO STEP4;
[STEP4]
考虑隶属于当前结点 pCurrentNode 的子实例集,对任一离散属性,属于不同概念例子的取值是否都相同?
是,则 pCurrentNode→nNodeType = NN;
调用 FTART 算法生成一个新的神经网络结点;
返回;
否则,GOTO STEP5;
[STEP5]
调用 BS 算法,生成 pCurrentNode→nAttr,pCurrentNode→nAttrValue,pCurrentNode→nConceptIndex;
生成当前结点的左子树和右子树;
返回;
[STEP6]
是否 newExam 拥有属性值对〈pCurrentNode→nAttr,pCurrentNode→nAttrValue〉?
是,则 pCurrentNode =pCurrentNode→pLpointer;
否则 pCurrentNode=pCurrentNode→pRpointer;
GOTO STEP1.

3 运行实例

3.1 学习结果概述

我们用 HLM 模型对台风路径分类、植物分类以及 Quinlan 的人群分类[12]等问题进行了学习,采用增量学习方式后效率明显提高.

在台风路径分类问题中,例子集共有 51 个例子,每个例子有 5 种离散属性、8 种连续属性. 增量学习前测得正确率为 93.76%. 采用增量学习方式增加 3 个例子后正确率达到 98.85%,耗时仅 6s. 如果不采用增量学习方式,重新生成判定树将耗时 24s.

在植物分类问题中,例子集共有 132 个例子,每个例子有 8 种离散属性、10 种连续属性. 增量学习前测得正确率为 88.74%. 采用增量学习方式增加 3 个例子后正确率达到 91.72%,耗时仅 7s. 而如果不采用增量学习方式,重新生成判定树将耗时 41s.

限于篇幅,在此仅对人群分类问题予以详细介绍.

3.2 人群分类问题的运行实例

首先,我们把背景知识输入到 HLM 中,这包括:

概念集{WHITE,BLACK,YELLOW},其含义分别为白种人、黑种人、黄种人.

离散属性集{hair,eye},其含义为头发的颜色和眼睛的颜色.

离散属性值集为 hair{blond,red,dark,gray};eye{blue,dark,brown}.

连续属性集{height,weight},其含义分别为身高、体重.

连续属性的取值范围为 height(150～220);weight(40～125).

训练实例集如表 2 所示,最初生成的混合型判定树如图 1 所示.

以增量学习方式增加训练例(gray,blue,188,79,white),得到图 2 所示的判定树.

再增加训练例(dark,dark,146,42,yellow),该实例将匹配到 FTART 结点 NN1. NN1 原有 7 个训练模式对,其网络结构如图 3 所示,图中白色神经元表示兴奋神经元. 进行增量学习后,NN1 有 8 个训练模式对,其网络结构如图 4 所示. 从图中可以看出,网络第 2 层增加了一个兴奋神经元.

在以增量学习方式增加以上两个实例之前测得正确率为 92.12%,增量学习后正确率

达到 95.44%,耗时仅 6s,而如果重新生成判定树则需耗时 39s.

表 2　人群分类实例集

index	hair	eye	height	weight	concept	index	hair	eye	height	weight	concept
1	blond	blue	189	80	white	17	gray	dark	175	95	black
2	blond	blue	199	100	white	18	gray	dark	185	100	black
3	red	brown	198	105	white	19	gray	dark	185	100	black
4	blond	brown	199	100	white	20	gray	dark	185	100	black
5	red	dark	220	125	white	21	dark	dark	155	50	yellow
6	red	dark	205	125	white	22	dark	dark	165	60	yellow
7	dark	brown	210	120	white	23	dark	dark	168	60	yellow
8	dark	dark	220	122.5	white	24	dark	dark	153	45	yellow
9	dark	dark	215	117.5	white	25	dark	dark	132	42.5	yellow
10	dark	dark	170	70	black	26	dark	dark	166	50	yellow
11	dark	dark	178	73	black	27	dark	brown	160	47.5	yellow
12	dark	dark	180	82.5	black	28	dark	brown	167	53.5	yellow
13	dark	dark	175	95	black	29	dark	brown	160	47.5	yellow
14	dark	dark	185	100	black	30	dark	brown	167	53.5	yellow
15	gray	dark	178	73	black	31	dark	brown	160	47.5	yellow
16	gray	dark	180	82.5	black	32	dark	brown	167	53.5	yellow

图1　增量学习前的混合型判定树

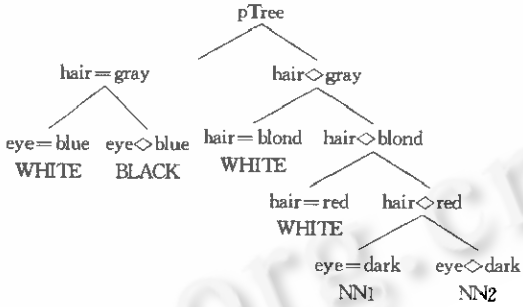

图2　增量学习后的混合型判定树

图3　增量学习前的NN1

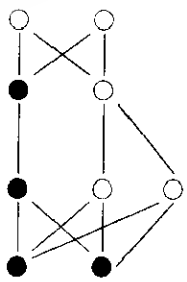

图4　增量学习后的NN1

4　结束语

混合型学习模型 HLM 通过神经网络与符号学习的有机结合,在机器学习的连续值输入、多概念获取等方面进行了有益尝试.其增量学习算法具有较强的增量学习能力,效率得

到了显著提高.该算法应用于江苏省气象台"台风路径预报系统"中,运行已半年多,对台风路径预报实例集进行学习,能提高台风路径预报的预报速度和准确率,取得了良好效果.

参考文献

1 Mooney R, Ourston D. A multistrategy approach to the theory refinement. In: Proceedings of the International Workshop on Machine Learning, Harper's Ferry, WV, 1991. 115～130.

2 Towell G G, Shavlik J W, Noordewier M O. Refinement of approximately correct domain theories by knowledge-based neural networks. In: Proceedings of the Eighth National Conference on Artificial Intelligence, Boston: AAAI Press, 1990. 861～866.

3 Utgoff P E. Perception trees: a case study in hybrid concept representations. In: Proceedings of the Seventh National Conference on Artificial Intelligence, St. Paul, MN: Morgan Kaufmann, 1988. 601～606.

4 Gallant S L. Connectionist expert systems. Communications of the ACM, 1988, **31**: 152～169.

5 陈兆乾,刘宏等.一种混合型多概念获取算法HMCAP及其应用.计算机学报,1996,**19**(**10**):753～761.

6 陈兆乾,周戎等.一种新的自适应谐振算法FTART.软件学报,1996,**7**(**8**):458～465.

7 Wasserman P D. Advanced methods in neural computing. 1993. 14～34.

8 Carpenter G A, Grossberg S, Reynolds J H. ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks, 1993, **4**: 565～588.

9 Moore B. ART 1 and pattern clustering. In: Proc. of 1988 Connectionist Models Summer School, 1989. 174～185.

10 Schlimmer J C, Fisher D A. Case study of incremental concept induction. In: Proceedings of the Fifth National Conference on Artificial Intelligence, Los Altos, CA: Morgan Kaufmann, 1986.

11 Utgoff P E. Incremental induction of decision trees. Machine Learning, 1989, **4**: 161～186.

12 Quinlan J R. Learning efficient classification procedures and their application to chess and games. In: Michalski R S, Carbonell J, Mitchell T M eds., Machine Learning: An Artificial Intelligence Approach, Los Altos, CA: Morgan Kauffmann, 1983.

THE INCREMENTAL LEARNING ALGORITHM IN HYBRID LEARNING MODEL HLM

CHEN Zhaoqian ZHOU Zhihua LI Hongbing XIE Junyuan

(*Department of Computer Science and Technology Nanjing University Nanjing* 210093)
(*State Key Laboratory for Novel Software Technology Nanjing University Nanjing* 210093)

Abstract The multi-concept acquisition algorithm HMCAP and the neural network algorithm FTART are integrated in hybrid learning model HLM, which can deal with multiple concepts and continuous attributes. In this paper, the incremental learning algorithm of HLM which based on the structure of hybrid binary decision tree and FTART network is proposed. It has the ability of adjusting old structure to improve learning accuracy by once learning instead of rebuilding the decision tree and the neural networks when the new examples were provided.

Key words Hybrid model, incremental learning, neural network.

Class number TP18