一种基于统计的神经网络规则抽取方法*

周志华, 何佳洲, 尹旭日, 陈兆乾

(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

E-mail: zhouzh@nju.edu.cn

http://aiake1.nju.edu.cn/~zhou

摘要: 从功能性观点出发,提出了一种基于统计的神经网络规则抽取方法.该方法利用统计技术对抽取出的规则进行评价,使其可以较好地覆盖示例空间.采用独特的连续属性处理方式,降低了离散化处理的主观性和复杂度.采用优先级规则形式,不仅使得规则表示简洁、紧凑,而且还免除了规则应用时所需要的一致性处理.该方法不依赖于具体的网络结构和训练算法,可以方便地应用于各种分类器型神经网络.实验表明,利用该方法可以抽取出可理解性好,简洁、紧凑,保真度高的符号规则.

关键词: 神经网络;规则抽取;机器学习;统计;聚类

中图法分类号: TP183 **文献标识码**: A

神经网络具有优越的非线性处理能力和泛化能力,在很多实际领域中都取得了传统符号学习机制所难以获得的效果.但神经网络方法存在一个固有的缺陷,即由于其获取的知识蕴涵在大量的连接权中难以理解,也难以为推理过程给出清晰的解释,这对使用者了解网络功能以及利用神经网络进行知识发现和知识精化等任务非常不利.这一缺陷已经严重地限制了神经网络的进一步发展.

如果能从神经网络中抽取易于理解的符号规则,就可以从根本上解决这一问题.目前,这方面的工作越来越受到重视,也取得了很多成果.本文从功能性观点出发,提出了一种从神经网络中抽取符号规则的通用方法,即"基于统计的产生-测试法"(statistics-based producing and testing,简称SPT).实验结果表明,SPT可以抽取出可理解性好,简洁、紧凑,保真度高的符号规则.

1 相关工作

对神经网络规则抽取的研究开始于20世纪80年代末.Gallant[1]基于推理强度对可用属性进行排序,从而构造出可以解释网络如何为某个给定事例产生结论的规则.Saito和Nakano[2]令网络输入逐渐改变,通过检查网络的激活度来构造候选规则集,并从中寻找有用的规则.Fu[3]搜索结点的扇入连接权,通过找出权值之和超过阈值的连接权子集来抽取规则.Towell和Shavlik[4]将相似权聚为等价类,从基于知识的神经网络中抽取MOFN规则.Sestito和Dillon[5]利用多层网络度量输入之间的接近程度,并利用单层抑制性网络度量输入、输出相关度,从而获得合取规则.Thrun[6]利用有效区间分析,根据网络的输入、输出行为抽取规则.Craven和Shavlik[7]将规则抽取视为一个学习任务,利用三用途Oracle调用和判定树抽取合取规则.Setiono[8]对隐层神经元激活值进行

* 收稿日期: 1999-06-25; 修改日期: 1999-12-03

基金项目: 国家自然科学基金资助项目(69875006);江苏省自然科学基金资助项目(BK99036)

作者简介: 周志华(1973-),男,江苏盐城人,博士,主要研究领域为神经网络,机器学习,数据挖掘;何佳洲(1966-),男,江苏镇江人,博士生,工程师,主要研究领域为神经网络,故障诊断;尹旭日(1964-),男,安徽蚌埠人,博士生,讲师,主要研究领域为数据挖掘,故障诊断;陈兆乾(1940-),女,安徽合肥人,教授,博士生导师,主要研究领域为机器学习,专家系统,神经网络.

聚类,并在需要时反复将网络分裂为子网.Benitez 等人[9]在一类神经网络与模糊规则系统之间建立等同性,从而使得网络可以用一组模糊规则加以解释.

上述的大多数方法都有遇到组合爆炸的危险,因此,它们多采用修剪网络[1,8]、聚类权值[4,5,8]或限制规则前件数[2,3]等方式降低组合复杂度.某些方法对网络结构[1,4]或训练算法[1]有特殊的要求,某些方法对网络输入[1~3]或神经元激活值[1~3,6]有特殊要求,还有一些方法只能抽取出复杂而难以理解的规则[6,9].因此,现阶段仍有必要对规则抽取问题进行深入的研究.

2 基于统计的产生-测试法 SPT

2.1 功能性观点

示例通常由输入模式和输出模式组成,如果用训练好的神经网络对示例进行判别,并将其判别结果作为输出模式,与原输入模式组成一个新的示例,则该示例就在一定程度上反映了网络在示例空间中该点上的响应特性.如果这种示例的数目足够多,并且比较均匀地覆盖整个示例空间,则从该示例集中抽取出的规则将具有与原神经网络相似的使用效果,即这些规则可以描述原网络的功能.这种功能性的观点就是 SPT 方法的出发点.

2.2 统计技术的引入

利用网络生成的示例集未必能完备地覆盖整个示例空间,这就使得抽取出的规则未必能有效地描述神经网络的泛化能力.为了解决这个问题,我们在 SPT 中引入了统计方法.

在规则抽取之前,抽取出的规则集显然为空.随着抽取过程的进行,规则数不断增加.当一条新的规则被抽取出来时,它并不直接进入规则集.算法将利用网络再产生一些新的示例,并利用这些示例对规则进行评价.只有在规则对网络的保真度,即规则集与神经网络的判别结果符合程度满足要求时,该规则才被接受.在生成新示例时,对规则集中所有规则前件所涉及的属性取固定值,其他属性在值域内取随机值,并利用网络产生相应的示例输出模式.这样,新示例将完全符合已抽取出的规则集,并满足当前规则的前件部分.

如果将已抽取出的规则所覆盖的示例空间称为已知空间,未覆盖部分称为未知空间,则新示例的生成过程相当于在未知空间中,通过对当前规则前件所未加以限制的部分进行扰动,以使示例尽可能分布在整个区域,从而判断出当前规则是否对整个未知空间都有效.只有以某一概率对整个未知空间有效的规则才会被接受,其结果将使未知空间进一步缩小.这样,随着规则抽取的进行,规则集将逐渐覆盖整个示例空间,即逐步逼近原神经网络的功能.

2.3 连续属性处理机制

从处理连续属性的神经网络中抽取规则是一个非常困难的课题,目前还没有较好的方法.如果不对连续属性进行离散化处理,以类似于回归树[10]的方式进行反复的区间切分,则会由于连续属性取值空间的广大而陷入组合爆炸.因此,适当的离散化处理是必要的.

一些研究者在规则抽取开始时对所有连续属性进行统一的离散化处理[7],这样虽然可以简化问题,但由于示例空间的内在分布特性未知,使得这种方法具有较大的主观性,存在很多缺陷.首先,并非所有的连续属性都需要进行离散化处理.在属性较多的实际问题中,很多属性并不会出现在最终的规则中,如果也对它们进行离散化处理,则将增加很多额外计算开销.其次,由于事先并不知道哪些属性具有较好的聚类性能,当规则抽取过程在选择规则前件时,如果同时面临几个连续属性,则难以进行取舍.第 3,各连续属性取值的分布很可能有不同的特性,其合适的聚类数也应该不

尽相同,如果一律将其离散化为相同数目的区间,则将极大地抹杀连续属性所包含的分类信息.

SPT 并不在规则抽取开始时进行离散化处理.该算法首先从离散属性中抽取规则,当离散属性不足以缩小未知空间时,就选择一个聚类效果最好的连续属性进行离散化,并将其作为一个新的离散属性用于规则抽取.由于每次只对一个连续属性进行处理,不仅可以挑出当前具有最佳聚类性能的属性,而且还能保证不会离散化不必要的属性.更重要的是,离散化区间数可以根据属性值在未知空间的分布情况来确定,不同的属性可以聚为不同数目的类,从而可以更灵活地反映实际分布的情况.此外,随着规则抽取过程的进行,未知空间越来越小,各属性的离散化复杂度也越来越小,这使算法在连续属性处理上的开销得以减少.

2.4 优先级规则表示形式

在 SPT 规则抽取过程中,由于不断地对连续属性进行离散化处理,规则中用到的属性数将不断增长.然而,这些属性取值的适用范围并不相同.更特殊的是,随着未知空间的缩小,一些属性将按不同的取值出现在不同的规则中,其适用范围也不相同.因此,普通的规则形式不适合于 SPT 规则抽取过程的特性.为了解决上述问题,SPT 采用了优先级规则,先抽取出的规则具有较高的优先级,对应于较大的未知空间;后抽取出的规则优先级较小,对应于较小的未知空间.在规则应用时,待判别示例将按优先级顺序从高到低与规则进行匹配.这样,规则与属性取值适用范围不同的问题就得到了解决.同时,这还使得规则集可以具有较紧凑的表示形式.这不仅是由于规则前件数较少,还由于每抽取一条规则,未知空间就缩小一次,下一次抽取的规则将不涉及已知空间,从而使得规则的冗余度较小.即为了描述相同的网络结构,SPT 所需的规则数目较少.这一性质在面对规模较大的神经网络时可以表现出较大的优势.此外,在使用普通的规则集时,通常都需要进行一致性检查,以免较特殊的规则被较一般的规则所取代.优先级规则就完全免除了这一要求,方便了规则的应用.

2.5 算法描述

(1) 利用训练好的神经网络生成一个规则抽取示例集 S;

(2) 如果 S 中存在某个离散输入属性组合,具有该组合的示例属于某一分类,则将该输入属性组合作为规则前件,相应的分类作为规则后件,建立一条规则(如果有多个这样的属性组合,则选择覆盖示例最多的一个);如果不存在这样的离散输入属性组合,则执行步骤(5);

(3) 利用神经网络生成 N 个新示例,如果规则集的保真度不满足要求,则拒绝步骤(2)中新建立的规则,执行步骤(2),否则就接受该规则;

(4) 从 S 中去掉新规则覆盖的示例,如果 S 为空,则算法结束,否则执行步骤(2);

(5) 如果还有未离散化的连续属性,则选择聚类效果最好的一个进行聚类处理,执行步骤(2);

(6) 对具有连续优先级并且后件相同的规则进行合并处理;

(7) 结束.

参数 N 视用户要求而定.如果对规则可靠性要求较高,则 N 应取较大的值.一般情况下,当 N 与 S 中的示例数大致相等时即可获得相当精确的规则.

3 实验结果与比较

3.1 实验结果概述

我们用 SPT 从处理 LED(low emitting diode)数字识别问题和人群分类问题的神经网络中抽

取规则.在LED数字识别问题中,我们将SPT与Sestito方法[5]进行了比较;在人群分类问题中,我们将SPT抽取的规则与原神经网络以及C4.5判定树[11]进行了比较.比较结果说明了SPT方法的有效性.为了说明SPT的通用性,我们用目前最常用的单隐层BP网络来解决这两个问题,并对训练好的BP网络进行规则抽取.

3.2 LED数字判别问题

LED通过7个位置标记来显示十进制数字0~9,该问题就是要根据相应于这7个位置标记的输入属性来判别LED所显示的数字.示例集中共有3000个示例,其生成方式与文献[5]中的相同,引入了10%的随机误差.我们建立的BP网络有7个输入神经元,分别对应于7个位置标记;有10个输出神经元,分别对应于十进制数字0~9;还有8个隐层神经元.输入层神经元的含义如图1所示.

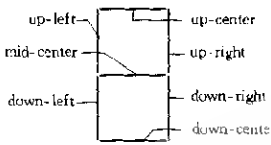


Fig.1 Meanings of input units in LED problem
图1 LED数字判别问题输入神经元含义

我们用SPT对神经网络进行规则抽取,参数N设置为与S中的示例数相同.抽取出的规则经整理后见表1,显然,这些规则都是正确的.

Table 1 Rules extracted via SPT in LED problem
表1 SPT方法为LED数字判别问题抽取的规则

Rule No.①	Rule②
1	up-center, up-left, up-right, mid-center, down-left, down-right, down-center→Eight
2	up-center, up-left, up-right, down-left, down-right, down-center→Zero
3	up-center, up-left, up-right, mid-center, down-right, down-center→Nine
4	up-left, up-right, mid-center, down-right→Four
5	up-center, up-right, mid-center, down-right, down-center→Three
6	up-center, up-right, down-right→Seven
7	up-right, down-right→One
8	up-center, up-left, mid-center, down-left, down-right, down-center→Six
9	up-center, up-left, mid-center, down-right, down-center→Five
10	up-center, up-right, mid-center, down-left, down-center→Two

①规则序号,②规则.

从表1可以看出,较一般的规则和较特殊的规则同时存在,如规则1和规则2.Sestito方法也抽取出了与表1类似的10条规则[5],但由于其规则之间没有联系,因此,在使用这些规则时必须进行一致性判断,以避免较特殊的规则在应该出现的时候被较一般的规则所代替.而SPT抽取出的规则是具有优先级的,不存在这个问题.因此,SPT抽取的规则具有更好的实用性.

3.3 人群分类问题

该问题的示例集见文献[12],每个示例都具有两个离散属性和两个连续属性.由于出现了连续属性,SPT需要进行离散化处理.这里,我们只用了最简单的K均值聚类法.SPT算法的参数N仍设置为与S中的示例数相同.

由于抽取出的规则精度难以直观判断,我们将原示例集[12]留作测试集,通过引入10%的随机误差,利用其构造出一个新的示例集用于神经网络训练,该示例集包含160个示例.我们分别训练了5个BP网络,其输入神经元数为4,输出神经元数为3,隐层神经元数分别为2,3,5,7,9.因为5次实验中训练出的网络不同,SPT抽取的规则也有轻微的变化.抽取出的一个典型规则集见表2.

Table 2 Rules extracted via SPT in human race problem

表 2 SPT 方法为人群分类问题抽取的规则

Rule No.①	Rule②
1	(hair=blond) ∨ (hair=red)→White
2	(hair=gray)→Black
3	(193.6<height)→White
4	(height<162.7)→Yellow
5	(65.3<weight)→Black
6	(weight<65.3)→Yellow

①规则序号,②规则.

在每次实验中我们都用测试集对 C4.5 判定树、原神经网络以及 SPT 从神经网络中抽取的规则集进行测试,其比较结果见表 3.其中"SPT 保真度"一列表示 SPT 规则集与原神经网络判别结果相同的测试例在测试集中所占的比例.

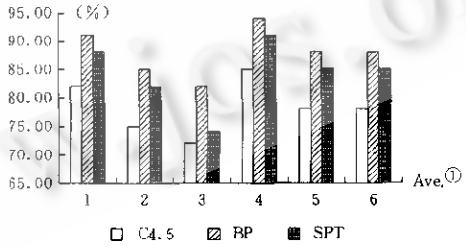
Table 3 Comparison of human race problem test results

表 3 人群分类问题测试结果比较

Exp. No.①	Test accuracy②			SPT fidelity③(%)
	C4.5 (%)	BP (%)	SPT rule (%)	
1	81.3	90.6	87.5	96.9
2	75.0	84.4	81.3	93.8
3	71.9	81.3	78.1	90.6
4	84.4	93.8	90.6	96.9
5	78.1	87.5	84.4	90.6
Ave.④	78.1	87.5	84.4	93.8

①实验序号,②测试精度,③SPT 保真度,④平均值.

从表 3 可以看出,神经网络模型具有较高的精度.这是由于提供给神经网络和 C4.5 判定树学习的训练例存在噪音,而神经网络具有较好的泛化能力和容噪性,因此能取得较好的效果.图 2 直观地对 3 种方法进行了比较.从图中可以看出,SPT 规则的测试精度与原神经网络测试精度相当接近,这充分说明了 SPT 方法的有效性.



①平均值.

图2 人群分类问题测试结果比较

Fig. 2 Comparison of human race test results

值得注意的是,在所进行的 5 次实验中,SPT 规则的测试精度都高于 C4.5 判定树.这与 Craven[7],Setiono[8]等人的观察结果一致,其原因是从神经网络中抽取的规则得益于神经网络的泛化能力,比判定树具有更好的预测性能.

此外,我们还用 SPT 算法从训练好的 FTART(field theory based adaptive resonance theory)[13],FTART2[14]网络中抽取规则,都取得了很好的效果,这充分显示出 SPT 具有较好的通用性.

4 结束语

神经网络已得到了广泛的应用,但其知识表示隐藏在大量连接权中,可理解性差,这已成为制约神经网络技术进一步发展的瓶颈.本文提出了一种基于统计的神经网络规则抽取方法,可以应用于各种执行分类学习任务的神经网络模型.实验结果表明,该方法能抽取可理解性好,简洁、紧凑,保真度高的符号规则集.我们已将其用于台风知识挖掘[15]技术之中,取得了很好的效果.

进一步的工作主要是提高 SPT 方法中连续属性离散化的处理性能,在高维非线性属性空间中也要能进行合适的聚类.此外,如果能根据连续属性之间聚类效果的相互影响程度,有联系地同时对多个连续属性进行离散化,不仅可以降低组合复杂度、减少计算开销、更好地用规则集逼近神经网络的性能,还可以发现数据之间更深层次的联系,挖掘出具有高度实用价值的知识.

References:

[1] Gallant, S. I. Connectionist expert systems. Communications of the ACM, 1988,31(2):152~169.

[2] Saito, K., Nakano, R. Medical diagnostic expert system based on PDP model. In: IEEE Neural Network Council ed. Proceedings of the IEEE International Conference on Neural Networks. New York: IEEE Press, 1988. 255~262.

[3] Fu, L. M. Rule learning by searching on adapted nets. In: AAAI ed. Proceedings of the 9th National Conference on Artificial Intelligence. Anaheim, CA: AAAI Press, 1991. 590~595.

[4] Towell, G. G., Shavlik, J. W. Extracting refined rules from knowledge-based neural networks. Machine Learning, 1993, 13(1):71~101.

[5] Sestito, S., Dillon, T. Knowledge acquisition of conjunctive rules using multilayered neural networks. International Journal of Intelligent Systems, 1993,8(7):779~805.

[6] Thrun, S. Extracting rules from artificial neural networks with distributed representations. In: Tesauro, G., Touretzky, D., Leen, T., eds. Advances in Neural Information Processing Systems, Vol 7. Cambridge, MA: MIT Press, 1995.

[7] Craven, M. W., Shavlik, J. W. Extracting tree-structured representations of trained networks. In: Touretzky, D., Mozer, M., Hasselmo, M., eds. Advances in Neural Information Processing Systems, Vol 8. Cambridge, MA: MIT Press, 1996. 24~30.

[8] Setiono, R. Extracting rules from neural networks by pruning and hidden-unit splitting. Neural Computation, 1997,9(1):205~225.

[9] Benitez, J. M., Castro, J. L., Requena, I. Are artificial neural networks black box? IEEE Transactions on Neural Networks, 1997,8(5):1156~1164.

[10] Breiman, L., Friedman, J., Olshen, R., *et al*. Classification and Regression Trees. New York: Chapman and Hall Press, 1984.

[11] Quinlan, J. R. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[12] Chen, Zhao-qian, Liu, Hong, Zhou, Rong, *et al*. A hybrid algorithm for multi-concept acquisition and its application. Chinese Journal of Computers, 1996,19(10):753~761 (in Chinese).

[13] Chen, Zhao-qian, Zhou, Rong, Liu, Hong, *et al*. A new adaptive resonance theory algorithm. Journal of Software, 1996, 7(8):458~465 (in Chinese).

[14] Zhou, Zhi-hua, Chen, Zhao-qian, Chen, Shi-fu. Field theory based adaptive resonance neural network classifier. Journal of Software, 2000,11(5):667~672 (in Chinese).

[15] Zhou, Zhi-hua, Chen, Shi-fu, Chen, Zhao-qian. Mining typhoon knowledge with neural networks. In: IEEE Computer Society ed. Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence. Los Alamitos, CA: IEEE Press, 1999. 325~326.

附中文参考文献:

[12] 陈兆乾,刘宏,周戎,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.

[13] 陈兆乾,周戎,刘宏,等.一种新的自适应谐振算法.软件学报,1996,7(8):458~465.

[14] 周志华,陈兆乾,陈世福.基于域理论的自适应谐振神经网络分类器.软件学报,2000,11(5):667~672.

A Statistics-Based Approach for Rule Extraction from Neural Networks*

ZHOU Zhi-hua, HE Jia-zhou, YIN Xu-ri, CHEN Zhao-qian

(*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210093, *China*)

E-mail: zhouzh@nju.edu.cn

http://aiake1.nju.edu.cn/~zhou

Abstract: In this paper, from the functional point of view, a statistics-based approach for rule extraction from trained neural networks is proposed. This approach introduces statistical technique to evaluate extracted rules so that the rule set could well cover the instance space. It deals with continuous attributes in a unique way so that the subjectivity and complexity of discretization are lowered. It adopts ordered rule representation so that not only the rules have concise appearance but also the consistency process could be released when the rules are used. Moreover, this approach is independent of the architecture and training algorithm so that it could be easily applied to diversified neural classifiers. Experimental results show that the symbolic rules extracted via this approach are comprehensible, compact, and with high fidelity.

Key words: neural network; rule extraction; machine learning; statistics; clustering

* Received June 25, 1999; accepted December 3, 1999

Supported by the National Natural Science Foundation of China under Grant No. 69875006; the Natural Science Foundation of Jiangsu Province of China under Grant No. BK99036