基于神经网络的多示例回归算法*

张敏灵, 周志华+

(南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

A Multi-Instance Regression Algorithm Based on Neural Network

ZHANG Min-Ling, ZHOU Zhi-Hua+

(National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: Phn: 86-25-3593163, Fax: 86-25-3300710, E-mail: zhouzh@nju.edu.cn

http://cs.nju.edu.cn/people/zhouzh/

Received 2002-09-29; Accepted 2003-03-04

Zhang ML, Zhou ZH. A multi-instance regression algorithm based on neural network. *Journal of Software*, 2003,14(7):1238~1242.

http://www.jos.org.cn/1000-9825/14/1238.htm

Abstract: Through employing a new error function capturing the nature of multi-instance learning, a neural network based multi-instance regression algorithm is presented in this paper. Experiments on benchmark data sets show that this algorithm deals well with the multi-instance regression problems.

Key words: multi-instance learning; multi-instance regression; machine learning; neural network; neural network ensemble

摘 要: 通过重新定义全局误差函数,提出了一种基于神经网络的多示例回归算法,并在基准数据集上对该算法进行了测试,取得了较好的效果.

关键词: 多示例学习;多示例回归;机器学习;神经网络;神经网络集成

中图法分类号: TP183 **文献标识码**: A

20 世纪 90 年代以来,根据从导师或环境获取的例子进行学习的示例学习被认为是最有希望的一种机器学习途径[1].目前,对示例学习的研究大致可分为 3 种学习框架(learning framework)[2],即监督学习、非监督学习和强化学习.1997年,Dietterich等人[3]通过在药物活性预测(drug activity prediction)问题方面的研究工作,提出了多示例学习(multi-instance learning)的概念.在多示例学习中,每个训练包(bag)由多个示例组成,示例没有概念标记,但包有一个概念标记.若包中至少有一个示例是正例,则该包被标记为正(positive);若包中所有示例都是反例,则该包被标记为反(negative).学习系统通过对多个包所组成的训练集进行学习,以尽可能正确地预测训练集之外的包的概念标记.由于多示例学习具有广阔的应用前景和独特的性质,属于以往机器学习研究的一个盲区,因此在国际机器学习界引起了极大的重视,被认为是与监督学习、非监督学习和强化学习并列的第 4 种示例学

* Supported by the National Natural Science Foundation of China under Grant No.60105004 (国家自然科学基金); the Natural Science Foundation of Jiangsu Province of China under Grant No.BK2001406 (江苏省自然科学基金)

第一作者简介: 张敏灵(1979－),男,浙江杭州人,硕士生,主要研究领域为数据挖掘,机器学习.

习框架[2].

Dietterich 等人[3]在提出多示例学习的概念时指出,探明决策树、神经网络以及其他一些常用的机器学习方法是否可以通过修改以用于多示例学习,并在可行的情况下设计出这些方法的多示例版本,是一个非常值得研究的课题;另一方面,一些研究者[4~7]指出,在许多情况下采用输出为实值的多示例回归算法将更有助于实际问题的解决,例如在药物活性预测问题中,如果能得到实值表示的输出,就可以表征出分子绑定的强弱,这对药物设计会有更大的帮助.针对上述两方面的问题,本文提出了一种基于神经网络的多示例回归算法 BP-MIR(BP for multi-instance regression).该算法通过采用特殊的误差函数对 BP 神经网络[8]进行了扩展,在基准数据集[7]上的实验表明,BP-MIR 能够有效地解决多示例回归问题.

本文首先介绍多示例学习的研究现状,然后给出 BP-MIR 算法的详细描述及其在基准数据集上的实验结果,最后对进一步的研究方向进行讨论.

1 研究现状

20 世纪 90 年代中期,Dietterich 等人[3]对药物活性预测问题进行了研究.其目的是让学习系统通过对已知适于或不适于制药的分子进行分析,以尽可能正确地预测某种新的分子是否适合制造这种药物.该问题的困难之处在于,每一个分子都有很多种可能的低能形状,只要该分子的某一种低能形状与期望的绑定区域(binding site)紧密耦合,该分子就适于制药,而生物化学家目前只知道哪些分子适于制药,并不知道具体的哪一种形状起到了决定性作用.为了解决上述问题,Dietterich 等人将每一个分子作为一个包,而将分子的每一种低能形状作为包中的一个示例,由此提出了多示例学习.在此基础上,他们将分子的低能形状通过属性-值对的形式表示出来,提出了 3 种 APR(axis-parallel rectangles)学习算法,这些算法都是通过对属性值进行合取,在属性空间中寻找合适的轴平行矩形.Dietterich 等人发现,iterated-discrim APR 算法在药物活性预测问题上取得了最好的效果,而直接将 C4.5 决策树、BP 神经网络等常用的监督学习算法用于解决多示例学习问题效果很不理想.由此可见,如果不考虑多示例学习本身的特点,将难以很好地完成此类学习任务.

作为一种新的学习框架,多示例学习受到了理论机器学习界的极大关注.Long 和 Tan[9]首先对 APR 在多示例学习框架下的 PAC 可学习性(PAC learnability)进行了研究.他们证明,如果包中的示例是独立的,且符合积分布(product distribution),那么 APR 在多示例学习框架下是 PAC 可学习的.在此基础上,他们提出了一种具有很高的多项式时间复杂度的理论算法.此后,Auer 等人[10]通过分析多示例学习框架下 APR 的可学习性与 DNF 公式(DNF formulas)可学习性之间的关系,证明了如果包中示例不是独立的,则在多示例学习框架下对 APR 进行学习是一个 NP 完全问题.与此同时,他们提出了一种改进的理论算法,其学习复杂度比 Long 和 Tan 的算法低得多,且不再要求包中示例符合积分布.Auer[11]还进一步地将该理论算法转变为一种可以用于解决实际问题的应用算法 MULTINST,通过在麝香分子数据集上的实验显示出该算法在药物活性预测问题上的可用性.1998 年,Blum 和 Kalai[12]指出多示例学习框架下的 PAC 学习可以转化为单边或双边随机分类噪音下的 PAC 学习,他们还借助于统计查询模型(statistical query model),得到了一个略优于 Auer 等人[10]的结果的理论算法.

在多示例学习的应用研究方面,最具影响力的是 1998 年由 Maron 和 Lozano-Pérez[13]提出的多样性密度(diverse density)方法.对于属性空间中的某一点,如果该点附近出现的正包数越多,而反包示例出现得越远,则该点的多样性密度越大.他们使用梯度法来寻找多样性密度的最大点.目前,多样性密度方法已经分别应用到股票选择[13]、从序列图像中学习人的简单描述[13]、自然场景分类[14]等领域.

除了多样性密度方法以外,还有许多针对多示例学习的应用算法.2000 年,Wang 和 Zucker[15]通过结合惰性学习(lazy learning)和 Hausdorff 距离,成功地对 k-近邻(k-nearest neighbor)方法进行了扩展,提出了 Bayesian-kNN 和 Citation-kNN 两种方法用以处理多示例学习问题.同年,Ruffo[16]给出了 C4.5 决策树的多示例版本 Relic,并将其成功地应用于数据挖掘领域.2001 年,Chevaleyre 和 Zucker[17]对决策树算法 ID3 以及规则学习算法 RIPPER 进行了扩展,得到了多示例决策树算法 ID3-MI 以及多示例规则学习算法 RIPPER-MI.此外,我们将神经网络引入了多示例分类问题,提出了一种多示例神经网络分类算法 BP-MIP[18].

在早期的多示例学习研究中,研究者们关注的主要是输出为离散值的多示例分类问题,但在药物活性预测

等许多应用领域,如果能采用实值表示的输出,则更有助于问题的解决.最近,一些研究者开始关注输出为实值的多示例回归问题.Ray 和 Page[4]首先提出了一种基于 EM(expectation maximization)方法的多示例回归算法,并且证明寻找多示例回归问题的精确解是一个 NP 完全问题.Dooly 等人[5]的研究结果肯定了上述结论,他们还进一步指出了多示例回归问题与 DNF 公式具有相同的学习难度.Zhang 和 Goldman[6]将 EM 方法和多样性密度方法相结合,给出了名为 EM-DD 的通用多示例回归算法.此外,Amar 等人[7]对 kNN 算法、Citation-kNN 算法和多样性密度方法进行了扩展,使其可用于解决多示例回归问题,并且还给出了一种生成具有一定实际物理含义的基准数据集的方法.

2 BP-MIR

假设训练集共由 N 个训练包 $\{B_1,B_2,\ldots,B_N\}$ 组成,其中,每个训练包 B_i 对应的实值标记为 L_i,并含有 M_i 个示例 $\{B_{i1},B_{i2},\ldots,B_{iMi}\}$.此外,包中的每个示例均为一个 p 维的属性值向量(记包 B_i 中的第 j 个示例为 $[B_{ij1},B_{ij2},\ldots,B_{ijp}]^{\mathrm{T}}$,T 代表向量转置).

在多示例学习问题中,学习算法已知包的实值标记,但却无法获得包中示例的实值标记.因此,我们利用训练包的实值标记,在包的层次上定义神经网络的全局误差函数如下:

$$E=\sum_{i=1}^{N}E_i, \tag{1}$$

其中 E_i 为包 B_i 对应的输出误差.

研究结果表明[7],包的实值输出是由包中示例的最大实值输出所决定的.因此,为了模拟上述规律,本文将包 B_i 的输出误差定义为

$$E_i=\frac{1}{2}\left(\max_{1\le j\le M_i}(o_{ij})-L_i\right)^2, \tag{2}$$

其中 o_{ij} 为示例 B_{ij} 对应的网络输出.

采用上述误差函数并结合基本的 BP 算法[8],本文提出了一种基于神经网络的多示例回归算法——BP-MIR(BP for multi-instance regression),其伪码如图 1 所示.

```
BP-MIR(Epochs, Threshold)
    Initialize neural network Net;
    for (epoch=1; epoch<=Epochs; epoch++)
        GlobalErr=0;   //Set the initial value of global error to be zero
        for (i=1; i<=N; i++)
            Compute the output error Ei of bag Bi according to Eq.(2);
            GlobalErr=GlobalErr+Ei;
            The weights in Net are modified according to Ei and the
            weight-updated rule of BP algorithm[8];
        end
        If (GlobalErr<=Threshold)
            return Net;
        end
    end
    return Net;
```

Fig.1 Pseudo code of BP-MIR algorithm

图 1 BP-MIR 算法的伪码表示

图 1 中的参数 *Epochs* 和 *Threshold* 分别为最大的训练轮数以及全局误差 E 的阈值.神经网络 *Net* 的输入神经元数等于示例的维数 p,并含有一层隐层神经元以及一个输出神经元,*Net* 中所有神经元的激活函数均采用 Sigmoid 函数:

$$f(x)=\frac{1}{1+\mathrm{e}^{-x}}. \tag{3}$$

3 实验结果及比较

2001 年,Amar 等人[7]根据计算分子之间“绑定耦合度(binding affinity)”的经验公式,给出了一种生成具有一定实际物理含义的基准数据集的方法.该方法在生成数据集时,可以灵活地控制包中示例的个数,示例的维数(属性数目),相关属性的数目与相关系数(权值)等参数.所有的数据集均采用 LJ-*r*.*f*.*s* 的命名规则,其中 *r* 代表相关属性的数目,*f* 代表示例的维数,*s* 代表属性相关系数的设置,如果 *s*=1,则表示存在 0 和 1 两种相关系数.

本文使用 BP-MIR 算法在 LJ-160.166.1-S, LJ-160.166.1, LJ-80.166.1-S 和 LJ-80.166.1 这 4 份数据集上进行了实验,对于每份数据集均采用 leave-one-out 的验证方法.其中,神经网络的输入神经元个数等于示例的属性数目(166),隐层神经元数为 80,输出层为单个神经元,学习率为 0.05,训练轮数为 1000.本文还将 BP-MIR 与 Diverse Density 以及 Citation-kNN 算法进行了比较,对于 Diverse Density 算法,其初始权值均为 0.1,而 Citation-kNN 的结果并未利用属性的权值信息,具体结果见表 1.其中,loss 代表算法的均方误差(包的实值标记均位于[0,1]区间内).此外,将输出大于 0.5 的包作为正包(概念标记为 1),输出小于 0.5 的包作为反包(概念标记为 0),本文还给出了算法的分类误差%err.

Table 1 Comparison of BP-MIR with Diverse Density and Citation-kNN

表 1 BP-MIR 与 Diverse Density 以及 Citation-kNN 算法的比较

Data set	BP-MIR		Diverse Density		Citation-kNN	
	%err	Loss	%err	Loss	%err	Loss
LJ-160.166.1-S	18.5	0.073 1	0.0	0.005 2	0.0	0.002 2
LJ-160.166.1	16.3	0.039 8	23.9	0.085 2	4.3	0.001 4
LJ-80.166.1-S	18.5	0.075 2	53.3	0.116	0.0	0.002 5
LJ-80.166.1	18.5	0.048 7	N/A	N/A	8.6	0.010 9

由表 1 可以看出,BP-MIR 能够有效地处理多示例回归问题,其结果优于 Diverse Density 方法,但不如 Citation-kNN 方法.值得注意的是,Amar 等人[7]指出,Diverse Density 与 Citation-kNN 两种方法对于示例中所含相关属性的数目相当敏感.例如,对于数据集 LJ-160.166.1 与数据集 LJ-80.166.1 而言,当示例中的相关属性数目由 160 降为 80 的时候,Citation-kNN 的均方误差由 0.0014 增至 0.0109,增幅近 800%,而 BP-MIR 的增幅仅为 22.4%.

此外,本文还初步地将神经网络集成[19]引入多示例回归问题的解决.对于数据集 LJ-80.166.1,本文采用 Bagging[20]方法生成了 4 份个体网络,并与表 1 中已有的 1 份个体网络构成了一个含有 5 份个体网络的神经网络集成.结果表明,该数据集的均方误差(采用简单平均的方式集成)由 0.0487 降至 0.0455,而分类误差(采用相对多数投票的方式集成)则由 18.48%降至 11.96%.

4 结束语

本文通过采用特殊的误差函数,给出了一种基于神经网络的多示例回归算法,并将神经网络集成初步地引入了多示例学习领域.在基准数据集上的实验结果表明,该方法取得了较好的效果.

进一步的工作主要包括如何调整神经网络拓扑结构以及算法的参数配置,使其具有更好的泛化能力.此外,如果能够采用合适的属性选择机制,那么就能标定示例中各个属性在学习过程中的重要程度,这对于提高算法的性能将有更大的帮助.

References:

[1] Mitchell TM. Machine Learning. New York: McGraw-Hill, 1997.

[2] Maron O. Learning from ambiguity [Ph.D. Thesis]. Department of Electrical Engineering and Computer Science, MIT, 1998.

[3] Dietterich TG, Lathrop RH, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 1997,89(1-2):31~71.

[4] Ray S, Page D. Multiple instance regression. In: Brodley CE, Danyluk AP, eds. Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001. 425~432.

[5] Dooly DR, Goldman SA, Kwek SS. Real-Valued multiple-instance learning with queries. In: Abe N, Khardon R, Zeugmann T, eds. Proceedings of the 12th International Conference on Algorithmic Learning Theory. Berlin: Springer-Verlag, 2001. 167~180.

[6] Zhang Q, Goldman SA. EM-DD: An improved multiple-instance learning technique. In: Dietterich TG, Becker S, Ghahramani Z, eds. Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2002. 1073~1080.

[7] Amar RA, Dooly DR, Goldman SA, Zhang Q. Multiple-Instance learning of real-valued data. In: Brodley CE, Danyluk AP, eds. Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001. 3~10. http://www.cs.wustl.edu/~sg/multi-inst-data.

[8] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature, 1986,323(9):533~536.

[9] Long PM, Tan L. PAC learning axis-aligned rectangles with respect to product distribution from multiple-instance examples. Machine Learning, 1998,30(1):7~21.

[10] Auer P, Long PM, Srinivasan A. Approximating hyper-rectangles: Learning and pseudo-random sets. Journal of Computer and System Sciences, 1998,57(3):376~388.

[11] Auer P. On learning from multi-instance examples: empirical evaluation of a theoretical approach. In: Fisher DH, ed. Proceedings of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997. 21~29.

[12] Blum A, Kalai A. A note on learning from multiple-instance examples. Machine Learning, 1998,30(1):23~29.

[13] Maron O, Lozano-Pérez T. A framework for multiple-instance learning. In: Jordan MI, Kearns MJ, Solla SA, eds. Advances in Neural Information Processing Systems 10. Cambridge: MIT Press, 1998. 570~576.

[14] Maron O, Ratan AL. Multiple-Instance learning for natural scene classification. In: Koller D, Fratkina R, eds. Proceedings of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1998. 341~349.

[15] Wang J, Zucker J-D. Solving the multiple-instance problem: A lazy learning approach. In: Langley P, ed. Proceedings of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000. 1119~1125.

[16] Ruffo G. Learning single and multiple instance decision trees for computer security applications [Ph.D. Thesis]. Torino: Department of Computer Science, University of Turin, 2000.

[17] Chevaleyre Y, Zucker J-D. Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem. In: Stroulia E, Matwin S, eds. Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence. Berlin: Springer-Verlag, 2001. 204~214.

[18] Zhou ZH, Zhang ML. Neural networks for multi-instance learning. In: Shi ZZ, He Q, eds. Proceedings of the International Conference on Intelligent Information Technology. Beijing: People's Post and Telecommunications Publishing House, 2002. 455~459.

[19] Zhou ZH, Chen SF. Neural network ensemble. Chinese Journal of Computers, 2002,25(1):1~8 (in Chinese with English abstract).

[20] Breiman L. Bagging predictors. Machine Learning, 1996,24(2):123~140.

附中文参考文献:

[19] 周志华,陈世福.神经网络集成.计算机学报,2002,25(1):1~8.