

DOI:10.3876/j.issn.1000-1980.2014.03.018

基于 Madaline 网络敏感性的主动学习算法研究

曾晓勤,赵倩倩,何嘉晟

(河海大学计算机与信息学院,江苏 南京 210098)

摘要:提出一种基于 Madaline 网络敏感性的主动学习算法。首先用部分样本训练 Madaline 网络,然后以 Madaline 网络输出对其输入在给定样本点附近变化的敏感性为尺度,主动从未参与训练的样本中挑选敏感性相对大的样本继续进行训练,循环反复这个过程直到满足训练要求为止。实验验证了该主动学习算法处理离散分类问题的有效性和可行性。

关键词: Madaline 网络;网络敏感性;主动学习

中图分类号: TP183 **文献标志码:** A **文章编号:** 1000-1980(2014)03-0278-05

An active learning algorithm based on sensitivity of Madaline network

ZENG Xiaoqin, ZHAO Qianqian, HE Jiasheng

(College of Computer and Information Technology Engineering, Hohai University, Nanjing 210098, China)

Abstract: This paper presents an active learning algorithm based on the Madaline network's output sensitivity due to its input variation near a given sample. First, a portion of samples are used to train a Madaline network. Then, based on the sensitivity of the network, some other samples with higher sensitivity are actively selected and added into the training data to continue the training of the network. This process is repeated until the training requirement is met. Experiments verified the effectiveness and feasibility of the proposed active learning algorithm in treating discrete classification problems.

Key words: Madaline network; network sensitivity; active learning

近年来,主动学习成为机器学习领域的研究热点。传统的监督学习算法多是被动学习,即只能被动接受指定的数据,要求训练样本中含有标注数据。然而,在很多实际问题中,如文本分类、图像检索、语音识别等往往含有大量的未标注数据,只有少量的标注数据,这使得监督学习算法在应用上存在很大的局限性^[1]。另外,对样本进行数据标注常常比较困难,甚至代价十分昂贵,而获取未标注数据的样本则相对容易。如针对基因序列的测试,标注一段基因序列需要进行代价昂贵的实验,相反,获取基因片段代价则相对小的多^[2]。在有大量标注数据的情况下,选择尽可能少的、有代表性的样本进行学习,将大幅度提高学习效率。

主动学习一般分为两部分:学习器和选择器。学习器可以是一个分类器,用于学习标注数据的样本,算法也多是监督学习中常见的算法。选择器用于选择未参加学习的、标注或未标注数据样本,是主动学习的核心。主动学习与传统被动学习最本质的区别是多了一个主动选择的过程,可以自主地选择最有“价值”的样本(对于未标注数据的,需交由专家进行标注),然后再带回学习。显然,对适用于少量标注数据与大量未标注数据共存情况的主动学习进行研究既具有学术研究的意义,也具有实际应用的价值。根据获得未标注数据样本方式的不同,可以将主动学习算法分为 2 种类型:基于流的主动学习^[3-5]和基于池的主动学习^[6-8]。

Madaline 网络是一种通用的前馈离散型多层网络。理论上,离散问题可作为连续问题的特例,所以可以用连续前向网络代替离散前向网络。但是,对于那些本质为离散,且连续技术难以适用的领域,如逻辑运算、分类以及聚类等,离散前向网络直接突显出简单和高效的优势。激活函数的不连续性使得离散前向网络不

收稿日期:2013-09-16

基金项目:国家自然科学基金(60971088)

作者简介:曾晓勤(1957—),男,上海人,教授,博士,主要从事计算机科学、计算智能等研究。E-mail:zxeng@hhu.edu.cn

能再使用成熟的 BP 算法^[9-10]。Winter 等^[11-12]对 Madaline 网络训练算法进行了研究,提出了著名的 MRII 算法。MRII 算法把对网络的训练转化为对每个 Adaline 的训练,以最小扰动原则为依据,通过不断的翻转与迭代找到最佳的权值。近些年,Zhong 等^[13]在 MRII 算法的基础上,利用 Madaline 网络的敏感性技术,对 Madaline 网络训练算法进行了进一步的研究和改进。

一般而言,神经网络敏感性是研究因网络参数的扰动而引起网络输出发生变化的情况^[14],需找出网络输出变化与网络参数扰动之间的依赖关系。若考虑参数仅为输入情况,由于不同输入样本的扰动可能引起输出的不同变化,所以针对样本扰动的网络敏感性可以用来衡量网络输出对于不同输入样本的敏感程度。以网络敏感性为尺度,不难给出如下 2 种解释和推论:(a)那些使得网络输出敏感性大的样本(称为敏感样本)离网络已建立起的判定边界一定相对较近,因而类似支持向量在支持向量机中扮演的角色,敏感样本值得挑选出来进行进一步精化训练;(b)那些使得网络输出敏感性小的样本(称为不敏感样本)离网络已建立的判定边界一定相对较远,如果不敏感样本与已参加训练的所有样本距离都较远,说明网络前期训练中缺失了这种样本所代表的信息,因而也需挑选出来进行进一步全面训练。所以,在处理分类问题的主动学习中,Madaline 网络敏感性可以作为一个合适的尺度,用来主动挑选尚未训练的样本参加训练,显然,策略是挑选那些对当前 Madaline 网络敏感性大或小的样本。

由以上所述可见,主动学习过程中,关键是如何从大量的未标注数据样本中挑选最有“价值”的样本进行学习。本文提出一种新的主动学习算法,即基于 Madaline 网络敏感性的主动学习算法。

1 Madaline 网络模型及符号约定

神经元 Adaline 是神经网络 Madaline 的基本构成单元,组成网络中的一个节点。

1.1 Adaline 模型及符号

图 1 为一个神经元 Adaline 的结构示意图。

输入向量为
$$\mathbf{X} = (x_1, x_2, \dots, x_n)^T \quad (x_i \in R, i = 1, 2, \dots, n) \quad (1)$$
权向量为
$$\mathbf{W} = (w_1, w_2, \dots, w_n)^T \quad (w_i \in R, i = 1, 2, \dots, n) \quad (2)$$

偏置为 w_0 , 激活函数为对称硬极限函数,即

$$f(a) = \begin{cases} 1 & a \geq 0 \\ -1 & a < 0 \end{cases} \quad (3)$$

式中: a ——净输入,可表示为 $a = \mathbf{W}^T \mathbf{X} + w_0 = \sum_{i=1}^n w_i x_i + w_0$, 网络的输出为

$$y = f(a) = f\left(\sum_{i=1}^n w_i x_i + w_0\right) = \begin{cases} 1 & \sum_{i=1}^n w_i x_i + w_0 \geq 0 \\ -1 & \sum_{i=1}^n w_i x_i + w_0 < 0 \end{cases} \quad (4)$$

就功能而言,神经元 Adaline 实现了一种从多维空间向一维空间的逻辑映射关系。

1.2 Madaline 网络模型及符号

Adaline 是 Madaline 网络的最简单形式,Madaline 网络可由 Adaline 根据一定的规则连接而成。图 2 是 Madaline 网络模型结构示意图, $\mathbf{X}^i$ 、 $\mathbf{Y}^i$ ($1 \leq i \leq L$) 分别表示网络第 i 层的输入和输出。由本文介绍可知, $\mathbf{Y}^{k-1} = \mathbf{X}^k$ ($2 \leq k \leq L$)。Madaline 网络第一层输入 $\mathbf{X}^1$ 是整个网络的输入,由样本的输入提供;最后一层输出 $\mathbf{Y}^L$ 作为整个网络的输出。为方便表达,用 $n^0 - n^1 - \dots - n^L$ 表示一个有 L 层的网络, n^0 为网络的输入维数,

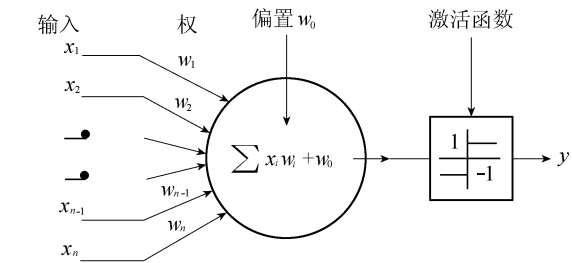


图 1 Adaline 结构示意图
Fig. 1 Sketch map of Adaline structure

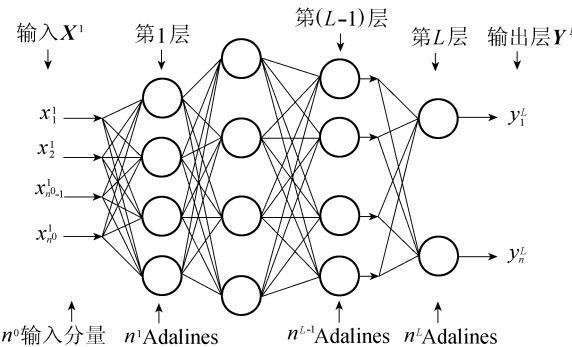


图 2 Madaline 网络模型
Fig. 2 Madaline network model

$n^i(1 \leq i \leq L)$ 为第 i 层上所包含的神经元个数。

从功能上看, Madaline 网络是在训练中通过一定的调整规则调整自身的参数(包括权值、偏置值或网络结构)来实现给定训练数据中输入与输出间的对应关系, 建立从 n^0 维空间到 n^L 维空间某种特定的映射关系。

2 网络敏感性定义及计算

2.1 Adaline 敏感性定义

影响 Adaline 输出变化的参数主要是权和输入。权决定输入与输出之间的映射关系, 因而它的扰动必然会对输入和输出关系产生影响, 进而使输出受到影响。权固定之后, 输入决定了输出, 输入扰动也势必会对输出产生影响。本文只考虑给定输入点的扰动对于输出的影响, 因此, Adaline 敏感性定义为: 因某一给定输入点的可能变化而导致输出的变化, 表示为

$$\Delta y = f((X_j + \Delta X)W + w_0) - f(X_jW + w_0)$$
 (5)

其中, 输入点 X_j 和权值 W 都是固定的。显然, 这个定义最自然和直接地表示了因输入点扰动 ΔX 而导致的输出变化 Δy 。

2.2 Madaline 网络敏感性定义

设 Madaline 网络输入为 $X_{n \times 1} \in R^n$, 输出为 $Y_{m \times 1} \in \{-1, 1\}^m$, 训练后的网络在输入和输出间建立的映射为 $G(X) = Y$, 又可表示为 $G(X) = (g_1(X), g_2(X), \dots, g_m(X))^T$ 。对 Madaline 网络敏感性进行研究, 已有文献^[15-16]针对不同的网络参数扰动情况进行了讨论。本文在 Adaline 敏感性定义的基础上, 定义 Madaline 网络敏感性为: 因网络在某一给定输入点的可能变化而导致网络输出变化的概率 P , 表示为

$$S = P_{\Delta X}(G(X_j + \Delta X) \neq G(X_j))$$
 (6)

其中, $\Delta X \in [-a, a]^n (a > 0)$ 视为一统计量, 是邻近输入点 X_j 周围的可能变化量。式(6)将 Madaline 网络敏感性定义为在给定的输入点 X_j 处, 由于 ΔX 而导致网络输出变化的概率。

2.3 Madaline 网络敏感性计算

由式(6)可得

$$S = P_{\Delta X}(G(X_j + \Delta X) \neq G(X_j)) = E_{\Delta X \in [-a, a]^n} \left(\frac{1}{2} |G(X_j + \Delta X) - G(X_j)| \right)$$
 (7)

式中: E ——期望值。则第 i 个分量的计算公式为

$$s_i = E_{\Delta X \in [-a, a]^n} \left(\frac{1}{2} |g_i(X_j + \Delta X) - g_i(X_j)| \right)$$
 (8)

对 Madaline 网络敏感性计算, 笔者给出一个简洁的近似计算方法, 基本思路是, 对输入扰动的每一个分量, 在连续区间 $[-a, a]^n$ 内离散取点, 然后计算输入扰动在所有可能离散点处的均值作为式(8)的近似值。显然, 离散点取的越多越精确, 但是计算复杂度将会呈指数级上升。

设第 i 个分量以步长 h 在区间 $[-a, a]$ 均匀取 $k = 2a/h + 1$ 个值, 则式(8)可近似计算为: 对于输入样本为 X_j , 输入变化范围为 $[-a, a]$ 的网络 $G() = (g_1(), g_2(), \dots, g_m())^T$, 其第 i 个输出分量敏感性的计算 $s_i(X_j, G, a)$ 为 $s_i = s_i + \frac{1}{2} |g_i(X_j + \Delta X) - g_i(X_j)|$, 则 $s_i = \frac{s_i}{k^n}$ 为第 i 个输出分量的敏感性, 其网络输出敏感性 $S(X_j, G, a)$ 即 $S = (s_1, s_2, \dots, s_m)^T$ 为网络在 X_j 的敏感性。

3 基于 Madaline 网络敏感性的主动学习算法

主动学习的基本思想是: (a) 根据初始的标注数据样本集生成基础分类器, 也就是利用监督学习算法对初始样本进行训练, 得到一个粗分类器; (b) 设计一个样本选择算法, 该算法根据一定标准对提供的未参与训练的样本(已标注或未标注数据样本)进行评价, 从中挑选一个或几个样本(未标注数据样本需交由专家进行标注)加入到训练数据集中, 继续对基础分类器进行训练。学习器和选择器重复交替工作, 逐步使分类器性能提高, 直到达到预定目标停止。

基于 Madaline 网络敏感性的主动学习算法以 Madaline 网络为学习器, 首先根据部分标注数据样本进行

学习,得到一个基础 Madaline 网络,然后以该 Madaline 网络在那些给定但未参与训练样本点的敏感性为评价尺度,挑选敏感性大的样本(对未标注数据样本需先进行标注)加入初始训练样本集中继续训练,循环往复该过程,直到最终满足学习器的性能目标。算法中之所以没有挑选敏感性小的样本,是因为在确定初始训练样本集时,可先把所有给定的训练样本进行聚类,挑选每个类中的样本作为初始训练样本集,这样就可避免感性值小的样本与参加训练样本的距离较远的情况。基于 Madaline 网络敏感性的主动学习算法具体步骤如下。

- a. 将训练样本集 X (其中有少量标注数据样本和大量未标注数据样本)聚类。
- b. 挑选每个类中的样本构成初始训练样本集 $\hat{X}$ (未标注样本需进行标注)。
- c. 循环执行以下操作直到满足终止条件:(a)用 $\hat{X}$ 训练分类器 G ;(b)对所有未参与初始训练的样本 $X_j \in (X-\hat{X})$,计算其敏感性值 $S(X_j, G, a)$;(c)按敏感性值递减的方法对样本排序;(d)选取前 k 个样本加入到 $\hat{X}$ 中(k 值可视具体应用而定)。
- d. 返回 G 。

4 实验结果及分析

实验以 VC++6.0 为实验平台进行,使用 UCI 数据库中的 Pima 数据集(两分类)和 Iris 数据集(三分类)。对于两分类,在构建分类器时输出层含有 1 个 Adaline 神经元;对于三分类,在构建分类器时输出层含有 2 个 Adaline 神经元。把每组数据集分成两部分:75% 的训练样本集和 25% 的测试样本集。在 X 中,首先将 X 进行聚类,然后挑选每个类中的样本构成 $\hat{X}$,这些 $\hat{X}$ 直接用于训练出一个粗分类器。根据主动学习的特性,在训练的过程中按照不同的挑选策略挑选参加训练的样本:(a)随机挑选策略,随机挑选未参与初始训练的样本加入到 $\hat{X}$ 中继续对基础分类器进行训练;(b)基于 Madaline 网络敏感性的主动学习挑选策略,用该策略挑选未参与初始训练的样本加入到 $\hat{X}$ 中继续对基础分类器进行训练。针对 Pima 数据集和 Iris 数据集,分别利用 2 种挑选策略进行了 5 组实验,每组实验针对相同的实验环境又做了 10 次,表 1 为每组实验的平均结果。

可以看出多数情况下本文提出的基于 Madaline 网络敏感性的主动学习算法比随机挑选的方法具有更高的准确率。

表 1 基于 Pima 数据集和 Iris 数据集 2 种挑选策略实验结果

Table 1 Experimental results of two methods based on Pima and Iris datasets

准确率/%			
Pima 数据集		Iris 数据集	
基于 Madaline 网络敏感性的主动学习挑选策略	随机挑选策略	基于 Madaline 网络敏感性的主动学习挑选策略	随机挑选策略
89.07	87.45	93.17	92.45
87.25	85.47	92.27	90.32
88.47	87.34	94.34	93.07
85.83	83.32	93.47	92.83
87.78	83.47	92.15	91.47

5 结 语

针对如何从大量的未标注数据样本中挑选最有价值的样本进行学习问题,本文提出了一种基于 Madaline 网络敏感性的主动学习算法。算法的核心是打造一种合适的 Madaline 网络敏感性尺度,有的放矢地选择需要的训练样本,从而实现主动学习的功能。实验结果验证了基于 Madaline 网络敏感性主动学习算法的有效性和可行性。

参考文献:

[1] 赵秋焕. 两种主动学习方法[D]. 石家庄:河北大学,2010.

[2] 崔宝今. 基于半监督和主动学习的蛋白质关系抽取研究[D]. 大连:大连理工大学,2008.

[3] FREUND Y, SEUNG H S, SHAMIR E, et al. Selective sampling using the query by committee algorithm[J]. Machine Learning, 1997, 28(2/3):133-168.

[4] DAGON I, ENGELSON S. Committee-based sample selection for probabilistic classifiers[C]//Proceedings of 12th Int'l Confon Machine Learning. San Francisco, CA: Morgan Kaufmann, 1995:150-157.

[5] ARGAMON-ENGLESON S,DAGAN I. Committee-based sample selection for probabilistic classifiers[J]. Journal of Artificial Intelligence Research,1999,11:335-360.

[6] LEWIS D D,GAIL W A. A sequential algorithm for training text classifiers;corrigendum and additional data[J]. Sigir Forum, 1995,29(2):13-19.

[7] MCCALLUM A K,NIGAM K. Employing EM and pool-based active learning for text classification[C]//Proceedings of the Fifteenth International Conference. Madison,WI,USA:Morgan Kaufmann Publishers, 1998:350-358.

[8] MUSLEA I,MINTON S,KNOBLOCK C A. Active learning with multiple view[J]. Journal of Artificial Intelligence Research, 2006,27:203-233.

[9] RUMELHART D E,HINTON G E,WILLIAMS R J. Learning representations by back propagation errors[J]. Nature,1986,323: 533-536.

[10] RUMELHART D E,MCCLELLAND J L. Parallel distributed processing:explorations in the microstructure of cognition,vol.1 [M]. Cambridge:MIT Press,1986.

[11] WINTER R,WIDROW B. Madaline rule II:a training algorithm for neural networks[C]//IEEE International Conference on Neural Networks. San Diego,CA,USA:IEEE Publishrs,1988:401-408.

[12] WINTER R. Madaline rule II:a new method for training networks for Adalines [D]. Stanford:Standford University,1989.

[13] ZHONG Shuiming,ZENG Xiaoqin,WU Shengli,et al. Sensitivity-based adaptive learning rules for BFNNs[J]. IEEE Transactions on Neural Networks and Learning Systems,2012,23(3):480-491.

[14] 曾晓勤,何嘉晟. 单隐层感知机神经网络对权扰动的敏感性计算[J]. 河海大学学报:自然科学版,2013,41(4):360-364. (ZENG Xiaoqin,HE Jiasheng. Computation the sensitivity of perceptrons with one hidden layer to weight perturbation[J]. Journal of Hohai University:Natural Sciences,2013,41(4):360-364. (in Chinese))

[15] WANG Yingfeng,ZENG Xiaoqin,YEUNG D,et al. Computation of Madalines' sensitivity to input and weight perturbations [J]. Neural Computation,2006,18(11):2854-2877.

[16] ZHONG Shuiming,ZENG Xiaoqin,LIU Huiyi,et al. Approximate computation of Madaline sensitivity based on discrete stochastic technique[J]. Science China:Information Science,2010,53(12):2399-2414.

· 简讯 ·

河海大学6个学科获得江苏省重点支持

在2014年5月8日召开的江苏省学科建设与研究生教育工作会议上,江苏省政府办公厅公布了江苏高校优势学科建设工程一期项目考核验收结果、二期项目立项学科和省重点序列学科名单。河海大学在一期工程中建设的水利工程、环境科学与工程、海岸带资源开发与安全学科群、土木工程安全与减灾学科群4个学科(群)考核验收结果均为“A”等;海洋科学、土木工程、水利工程、环境科学与工程4个学科入选二期工程,领域软件工程、技术经济管理2个学科入选省重点序列学科。

江苏高校优势学科建设工程是江苏省委、省政府颁布实施的《江苏省中长期教育改革和规划纲要(2010—2020年)》中的重要举措之一,旨在通过加强能力建设、人才培养和科研创新,建设一批高水平优势学科,全面提升江苏高等教育的水平和核心竞争力,使江苏高校学科整体实力在全国处于领先地位,部分学科进入国际前列。工程由省财政安排专项建设资金。

(本刊编辑部供稿)