基于机器学习统计思想实现多重线性回归分析

谷恒明[1],胡良平[1,2*]

(1.军事医学科学院生物医学统计学咨询中心,北京 100850;
2.世界中医药学会联合会临床科研统计学专业委员会,北京 100029
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍基于机器学习统计思想实现多重线性回归分析的方法。首先,简要回答了几个基本问题:什么是机器学习、它能解决的统计学问题及其具体方法;进一步,还粗略地介绍了 BP 神经网络回归分析方法的基本思路;最后,通过一个实例详细呈现了如何基于 R 软件实现 BP 神经网络回归分析的全过程。

【关键词】 回归分析;机器学习;误差反向传播神经网络;R 软件;输入节点;隐节点;输出节点
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.01.004

Realization of a multiple linear regression analysis based on the machine learning statistical thought

Gu Hengming[1], *Hu Liangping*[1,2*]

(1. *Consulting Center of Biomedical Statistics*, *Academy of Military Medical Sciences*, *Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies*, *Beijing* 100029, *China*
* *Corresponding author*: *Hu Liangping*, *E-mail*: *lphu812@sina.com*)

【**Abstract**】 The purpose of this paper was to introduce a method based on the machine learning statistics to implement a multiple linear regression analysis. First of all, it answered several basic questions briefly, such as what machine learning is, what statistical problems it can solve, what are its specific methods. Further, the brief introduction of the basic idea of BP neural networks regression analysis was presented. Finally, an example was given to show how to implement the whole process of BP neural networks regression analysis based on R software.

【**Keywords**】 Regression analysis; Machine learning; Error back propagation neural networks; R software; Input node; Hidden node; Output node

1 机器学习回归分析方法概述

1.1 何为机器学习

机器学习脱胎于人工智能[1],后者使“那些对人类智力来说非常困难、但对计算机来说相对简单的问题得到迅速解决”。人工智能的真正挑战在于解决那些对人来说很容易执行、但很难形式化描述的任务[2]。谁有能力“让计算机从经验获取知识,可以避免由人类来给计算机形式化地指定它需要的所有知识”?机器学习或深度学习具有这个能力。它致力于研究如何通过计算的手段,能够根据已有的输入数据进行学习,利用经验来改善系统自身的性能。具体地说,它是关于在计算机上从数据中产生“模型”的算法,即“学习算法”(learning algorithm)[3-4]的一门学问。

1.2 机器学习方法能解决的统计学问题

不难想象,对于统计表达与描述、基于检验统计量进行假设检验和相关性分析等问题,一般只需要采用经典统计方法、偶尔采用贝叶斯统计方法就可解决。而对于回归分析、判别分析和样品聚类分析等问题,虽然经典统计和贝叶斯统计常常也可以处理得令人比较满意,但若采取适当的机器学习方法(注意:具体的方法有很多种,如决策树、支持向量机、神经网络、集成学习和随机森林等[5]),常能产生意想不到的结果,尤其是在对资料前提条件的“零要求”和结果具有极高精准度等方面,更显示出极大的优越性。

前已述及,能够进行回归分析的机器学习方法有很多种,因篇幅所限,本文仅介绍基于 BP 神经网络方法实现多重线性回归分析的做法。

1.3 BP 神经网络回归分析方法简介

BP 神经网络是一种基于有监督的学习、使用非线性的可导函数作为传递函数的前馈神经网络,在 1986 年由 Rumlhart 和 McClelland 为首的科学家小组提出,是一种按误差逆传播算法训练的多层前馈网络,是目前应用最广泛的神经网络模型之一。BP 网络能学习和存贮大量的输入-输出模式映射关系,而无需事前揭示描述这种映射关系的数学方程。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

它的学习规则是使用最速下降法,通过反向传播来不断调整网络的权值和阈值,使网络的误差平方和最小。BP 神经网络模型拓扑结构包括输入层(input)、隐层(hide layer)和输出层(output layer)。

基本 BP 算法包括信号的前向传播和误差的反向传播两个过程。即计算误差输出时按从输入到输出的方向进行,而调整权值和阈值则从输出到输入的方向进行。正向传播时,输入信号通过隐含层作用于输出节点,经过非线性变换,产生输出信号,若实际输出与期望输出不相符,则转入误差的反向传播过程。误差反传是将输出误差通过隐含层向输入层逐层反传,并将误差分摊给各层所有单元,以从各层获得的误差信号作为调整各单元权值的依据。通过调整输入节点与隐层节点的联接强度和隐层节点与输出节点的联接强度以及阈值,使误差沿梯度方向下降,经过反复学习训练,确定与最小误差相对应的网络参数(权值和阈值),训练即告停止。一般情况下,隐层越多,误差越小,但是相应的模型越复杂。

BP 神经网络在网络理论和性能方面已经比较成熟。其突出优点是具有很强的非线性映射能力和柔性的网络结构。可根据具体情况设定网络的中间层及层数,并且随着结构的差异其性能也有所不同。但是 BP 神经网络也存在一些缺陷:①学习速度慢,即使是一个简单的问题,一般也需要几百次甚至上千次的学习才能收敛;②容易陷入局部极小值;③网络层数、神经元个数的选择没有相应的理论指导;④网络推广能力有限。

2 实例及机器学习回归分析的实现

2.1 问题与数据

【例 1】26 例糖尿病患者的血清总胆固醇(X_1)、甘油三酯(X_2)、空腹胰岛素(X_3)、糖化血红蛋白(X_4)、空腹血糖(Y)的测量值列于表 1,试基于机器学习统计思想(具体方法为“BP 神经网络”)建立血糖与其他几项指标间的多重线性回归方程,并完成其他有关的任务。

表 1 26 例糖尿病患者血样中有关指标的测定结果

i	X_1	X_2	X_3	X_4	Y
1	5.68	1.9	4.53	8.2	11.2
2	3.97	1.64	7.32	6.9	8.8
3	6.02	3.56	6.95	10.8	12.3
…	…	…	…	…	…
25	11.54	10.89	1.2	10.5	20
26	3.84	1.2	6.54	9.6	10.4

注:详细数据见本期第一篇文章《多重线性回归分析的核心内容与关键技术概述》

2.2 采用 BP 神经网络分析方法建模

将表 1 中的 26 行 5 列数据(不包括第 1 列编号,保留各列的变量名)输入计算机,用文本格式存储,取名为“空腹血糖与血脂.txt”。将其存储在 G 盘文件夹名为“studyr”的文件夹内。采用 R 软件包中的 nnet()函数来实现,所需要的 R 程序[设程序名为:BP 神经网络方法实现空腹血糖依赖三项血脂指标(注:X_1 无统计学意义,不参与建模)的多重线性回归分析.txt]如下:

```
install.packages("nnet")                    #安装实现 BP 神经网络计算的子程序包 nnet
library(nnet)                               #加载子程序包 nnet
setwd("G://studyr/")                        #设置路径为"G://studyr/"
data <- read.table("空腹血糖与血脂.txt",header=TRUE)
                                            #data 中的数据为 26 行 5 列
data1 <- data[,-1]                          #删除 data 中的第 1 列(即 X1)后为 data1
y <- data1[,4]/max(data1[,4])               #data1 中第 4 列除以该列最大值赋值给 y,y 为标准化结果
x <- data1[,-4]                             #data1 中的前 3 列(即 X2、X3、X4)赋值给向量 x
set.seed(1101)                              #设置随机数种子为 1101
                                            #每次修改 size(隐节点个数)的值,退出 R 运行环境;
                                            #再进入,结果具有重现性。
                                            #下面是基于 3-5-1BP 神经网络模型创建回归模型
(model1=nnet(x,y,size=5,entropy=TRUE,decay=0.1))
                                            #每次修改 size(隐节点个数)的
```

```
                                    值,退出 R 运行环境;
                                    #再进入,结果具有重现性。
                                    #设置 5 个隐节点,权重衰减速度的最小值为 0.1
summary(model1)                     #输出所创建的 BP 神经网络回归模型
pred1 = predict(model1,data1)       #计算标准化 y 的预测值
pred2 = pred1 * max(data1[,4])      #计算原始数据 y 的预测值
M <- sum((data1[,4] - mean(data1[,4]))^2)
                                    #计算原始数据 y 的离均差平方和
NMSE <- sum((data1[,4] - pred2)^2)/M
NMSE                                #输出标准化均方误差的数值
## data1[,4];pred2                  #删除语句前两个井号可输出观测 y 与预测 y 的数值
```

【说明】在上面的程序中,各语句后以“#”开头的就是该语句的解释信息。

2.3 BP 神经网络回归分析的结果

通常,采用一个隐藏层的 BP 神经网络,即 I - J - 1 BP 神经网络模型,其中,“I”为输入节点个数(具有统计学意义的自变量的个数);“J”为隐藏层中隐节点的个数;最后的“1”为输出节点个数(即定量结果变量的个数,回归分析时通常为 1;若是判别分析,通常为类别的个数)。

在上面的程序中,I = 3(3 个具有统计学意义的自变量,即 X_2、X_3 和 X_4)、J = 5(设置了 5 个隐节点)、1(代表只有 1 个输出节点,即定量的结果变量只有一个),故被称为“3 - 5 - 1BP 神经网络模型”。此模型的主要输出结果如下:

```
a 3-5-1 network with 26 weights
options were - entropy fitting dec
ay=0.1a 3-5-1 network with 26 weights
options were - entropy fitting decay=0.1
b->h1 i1->h1 i2->h1 i3->h1
 -0.07 0.17 -0.05 0.08
b->h2 i1->h2 i2->h2 i3->h2
 -0.10 0.16 -0.34 0.11
b->h3 i1->h3 i2->h3 i3->h3
 -0.07 0.17 -0.05 0.08
b->h4 i1->h4 i2->h4 i3->h4
 0.12 -0.15 0.42 -0.12
b->h5 i1->h5 i2->h5 i3->h5
 -0.10 0.16 -0.34 0.11
b->o h1->o h2->o h3->o h4->oh5
->o
 -0.05 0.36 0.48 0.36 -0.59 0.48
[1] 0.4166974
```

【说明】具有 26 个权重系数的 3 - 5 - 1 BP 神经网络模型,其含义如下:

三个自变量(被称为 3 个输入节点)分别用 i1、i2 和 i3 表示;“b”代表模型中的常数项;5 个隐节点(其本质就相当于因子分析中的“隐变量”,也可以将其视为“中间变量”,即它们是输入变量的结果变量、是输出变量的原因变量)分别用 h1 - h5 表示;输出节点(即定量的结果变量 y)用“o”表示,即“Output”之义。

在 3 - 5 - 1 BP 神经网络模型中,有(3 + 1) × 5 + (5 + 1) × 1 = 26 个权重系数。这个算式中的数字分别代表什么?“(3 + 1)”代表“三个自变量加一个常数项”共 4 项,每一项都要与 5 个隐节点相连接,故需要乘以 5;“(5 + 1)”代表 5 个隐节点加一个常数项共 6 项,每一项都要与一个输出节点相连接,故需要乘以 1。

任何一个“隐节点”与“输入节点(包括一个常数项)”之间都是通过“一个 Logistic 曲线”连接起来的,例如:

隐节点 h1 就是按下面的公式与四个输入节点连接起来的:

$$h1 = \exp(-0.07 + 0.17i1 - 0.05i2 + 0.08i3) / [1 + \exp(-0.07 + 0.17i1 - 0.05i2 + 0.08i3)]$$

同理,可以写出其他四个隐节点的表达式:

$$h2 = \exp(-0.10 + 0.16i1 - 0.34i2 + 0.11i3) / [1 + \exp(-0.10 + 0.16i1 - 0.34i2 + 0.11i3)]$$

$$h3 = \exp(-0.07 + 0.17i1 - 0.05i2 + 0.08i3) / [1 + \exp(-0.07 + 0.17i1 - 0.05i2 + 0.08i3)]$$

$$h4 = \exp(0.12 - 0.15i1 + 0.42i2 - 0.12i3) / [1 + \exp(0.12 - 0.15i1 + 0.42i2 - 0.12i3)]$$

$$h5 = \exp(-0.10 + 0.16i1 - 0.34i2 + 0.11i3) / [1 + \exp(-0.10 + 0.16i1 - 0.34i2 + 0.11i3)]$$

最后,可以写出一个输出节点的表达式:

$$0 = A/(1 + A) \quad (1)$$

其中,$A = \exp(-0.05 + 0.36h1 + 0.48h2 +$

0.36h3 -0.59h4 +0.48h5)

将上面的 h1 -h5 五个表达式代入公式(1),就可呈现用三个自变量表达结果变量的计算公式,其计算结果“o”为标准化后的预测值 y,即上面程序中的“Pred1”;将其乘以结果变量 y 的最大值,就将其还原为原始的结果变量的预测值了,即上面程序中的“Pred2”。

3 -5 -1 BP 神经网络模型得到的标准化均方误差 NMSE =0.416697。

在给定随机数和“I”数目的前提下,BP 神经网络拟合回归模型的计算结果的精确度可采用“标准化均方误差(NMSE)”的大小来度量。通常,随着隐节点的数目增大,NMSE 会逐渐变小。在本例中,分别取隐节点数目为:1、2、…、10、20、30、40、50、100、110、120、130、140、150 共 20 种情况时,对应的 NMSE 见表 2。

表 2 取 20 种不同数目的隐节点 BP 神经网络回归分析所产生的 NMSE 数值

隐节点数目	NMSE 的取值	隐节点数目	NMSE 的取值
1	0.477747	20	0.341086
2	0.454795	30	0.325781
3	0.433157	40	0.312347
4	0.417180	50	0.307483
5	0.416697	100	0.300271
6	0.401736	110	0.300449
7	0.388820	120	0.299277
8	0.388365	130	0.298621
9	0.375884	140	0.298564
10	0.376925	150	0.298454

结合本期前两篇文章中的计算结果和本文前面的计算结果可知:从模型的简练程度上看,BP 神经网络回归分析的模型很复杂,而传统统计和贝叶斯统计回归分析模型相对简练得多;从标准化均方误差(NMSE)大小上看,BP 神经网络回归分析的 NMSE 可以达到很小的程度,而很难使传统统计和贝叶斯统计回归分析模型的 NMSE 的数值降低很多。因此,基于 BP 神经网络回归分析与传统统计和贝叶斯统计回归分析之间的拟合效果不具有可比性。

3 小 结

当数据本身质量较好时,经典统计建模方法所要求的假定条件基本上能够得到满足,此时,经典统计建模的效果就比较好;如果有足够的先验信息,采用经典多重线性回归分析和贝叶斯回归分析的结果相差不大;BP 神经网络可以获得很精确的计算结果和精准的预测结果,但不便写出对应的回归模型。值得一提的是:只要结果变量为计量变量,希望研究其如何依赖自变量变化而变化的依赖关系时,选用多重线性回归分析比选用单因素差异性分析更合适。例如文献[6 -9]。

参考文献

[1] 郑捷. 机器学习算法原理与编程实践[M]. 北京:电子工业出版社, 2015:1 -55.

[2] 伊恩· 古德费洛, 约书亚· 本古奥, 亚伦· 库维尔. 深度学习[M]. 赵申剑, 黎彧君, 符天凡, 等译. 北京:人民邮电出版社, 2017:1 -18.

[3] 沙伊· 沙莱夫 -施瓦茨, 沙伊· 本 -戴维. 深入理解机器学习:从原理到算法[M]. 张文生译. 北京:机械工业出版社[M]. 北京:机械工业出版社, 2017:1 -9.

[4] 周知华. 机器学习[M]. 北京:清华大学出版社, 2016:1 -22.

[5] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 3 版. 北京:中国人民大学出版社, 2015:41 -56.

[6] 任传波, 姜季妍, 董黎明. 青少年抑郁障碍患者心理社会学特征[J]. 四川精神卫生, 2017, 30(5):455 -457.

[7] 徐华丽, 孙崇勇, 高悦. 大学生人格特质对手机成瘾倾向的影响[J]. 四川精神卫生, 2017, 30(5):458 -462.

[8] 李青青, 张倩. 医学院校大学生情绪状态与学业拖延的关系[J]. 四川精神卫生, 2017, 30(5):463 -465.

[9] 魏国英, 曾丽娟, 周桂成, 等. 精神科护士职业倦怠与工作压力的相关性[J]. 四川精神卫生, 2017, 30(5):466 -469.

(收稿日期:2018 -01 -29)

(本文编辑:陈 霞)