

# 机器学习在 GDP 预测分析中的应用研究

孙昊<sup>1</sup>, 张琦<sup>2</sup>, 许勇<sup>1,3</sup>

(1. 安徽师范大学 数学计算机科学学院, 安徽 芜湖 241000;

2. 西北大学 软件学院, 陕西 西安 710002;

3. 东南大学 计算机科学与工程系, 江苏 南京 210096)

**摘 要:**应用机器学习思想对 GDP 数据进行分析, 使用遗传算法优化 BP 人工神经网络建立 GDP 数据分析模型并进行预测, 带回验证表明模型具有较高精度。在机器已学得数据规律后, 利用 Sestito 和 Dillon 提出的 SD 算法, 对习得知识后的模型进行知识获取的分析, 得出一些由机器学习过程而获得的有意义的结论。这种分析方法可以广泛应用到如人口、经济等复杂系统的预测和分析中, 分析出相关因子对结果的影响程度, 为决策提供第三方的客观依据, 具有很强的推广性和实用性。

**关键词:**遗传算法; 人工神经网络; 规则抽取; GDP; SD 算法

**中图分类号:** TP181

**文献标识码:** A

**文章编号:** 1673-629X(2009)02-0227-03

## Study on Prediction and Analysis of GDP Based on Machine Learning

SUN Hao<sup>1</sup>, ZHANG Qi<sup>2</sup>, XU Yong<sup>1,3</sup>

(1. College of Mathematics and Computer Sciences, Wuhu Normal University, Wuhu 241000, China;

2. School of Software, Northwest University, Xi'an 710002, China;

3. Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

**Abstract:** In this paper, machine learning concept is applied to analyze GDP data. A BP neural network improved by genetic algorithm is applied to set up the predictive model then it is used to predict. When prove the model by giving the previous data, it showed a high accuracy. Then it is analyzed by SD algorithm put forward by Sestito and Dillon to extract the knowledge embedded within trained artificial neural network. The result gained by the process of machine learning is significative. This method can be widely applied to prediction and analysis of some complex system like population and economy, by providing the decision-maker the third-party suggestion, which has practicability.

**Key words:** genetic algorithm; artificial neural network; rule extraction; GDP; SD algorithm

## 0 引言

GDP 预测是一个研究的热点, 重要性不言而喻。而其较复杂的内在变动机理也增大了对 GDP 数据进行分析的难度。如何由 GDP 数据, 分析出增长规律并且提供有效信息以使决策者更好地对 GDP 增长进行预测、调控, 以使经济平稳、健康发展, 是一个有现实意义的问题。

人类在分析此类复杂数据的变化发展规律时, 通

过自己的经验、逻辑, 建立了一套基于已有数据的分析机制。可以说, 这套分析机制, 是人类大脑复杂学习过程的结果。然而人类大脑处理数据能力有限, 且主观性较强, 具有无法克服的缺点。而随着机器学习理论的发展, 人们可以通过第三方的智慧——计算机对此类数据进行客观、完整的分析。此类思想即, 通过计算机算法模拟智慧过程, 搭建一个具有学习机能的模型, 在模型的学习过程中对数据之间的关系、发展趋势等作出有限、定量的分析, 作为人们分析决策时的参考。

在机器学习方面, BP 人工神经网络是应用最为广泛的, 文中即以此为理论基础, 提供一种通过机器学习的过程对 GDP 数据进行分析的方法。利用 SD 算法对机器学习的习得结果进行分析, 得到 GDP 与影响因子的相关系数, 并讨论这样的第三方处理的实际意义。

收稿日期: 2008-05-12

基金项目: 安徽省自然科学基金重点资助项目(2005kj009zd)

作者简介: 孙昊(1987-), 男, 安徽蚌埠人, 研究方向为自然计算、人工智能、数据挖掘; 许勇, 博士, 教授, 硕士生导师, 研究方向为计算机网络、网络安全。

### 1 BP 人工神经网络与不足

人工神经网络,即人工搭建的模仿智慧生物大脑神经活动的仿真网络,其具有一定的智慧特征。在 1943 年 McCulloch 与 Pitts 提出第一个人工神经网络模型<sup>[1]</sup>以来,BP(Back Propagation)网络<sup>[2,3]</sup>是人工神经网络的重要一种,是应用最广泛、最成功的网络结构。BP 网络是一种多层前向神经网络,它由输入层、隐层、输出层组成,其中隐层可为单层或多层(见图 1)。网络采用有教师的训练,先进行正向传播,样本数据从输入层通过隐层神经元逐层进行处理,最终传向输出层。再使用反向传播算法,通过计算各层的实际输出与目标输出的差值,把误差信号按照原来的通路反向传回,并修改各层神经元的权值,以求误差信号趋于最小。致此,BP 神经网络完成学习过程。

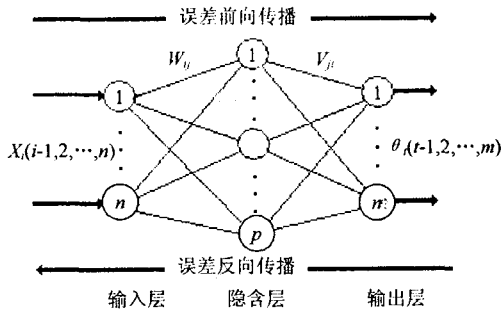


图 1 三层 BP 网络

利用 BP 网络的学习特性,可以把样本数据看成是一个时间序列,假定时间序列  $x_i = \{x_i \mid x_i \in R, i = 1, 2, \dots, L\}$ ,采用序列的前  $N$  个时刻的数据为滑动窗,利用 BP 神经网络将其映射为后  $M$  个时间的值。如果可以把数据分为  $K$  个长度为  $N + M$  的、有一定重叠的数据段,每一个数据段就可以看作一次预测实例。这些数据段的前部集合可以看作一个空间  $R_N$ ,如果可以将  $R_N$  合理映射到另一个空间  $R_M$  中,就有办法预知将来的  $M$  个数据,即:

$$x_{(n+k)} = F(x_{(n)}, x_{(n-1)}, \dots, x_{(n-m+1)})$$

如果不仅考虑已有数据对未来影响,还需要考虑其他因子对 GDP 的影响,则可以将其他因子也作为网络输入,由 BP 网络一并习得。

BP 网络的拟合能力强,具有优良的非线性逼近特性、大量的并行分布结构以及学习和归纳能力,但是它的不足也是非常明显的:收敛速度慢,容易陷入局部极小。针对这些缺点,可以用遗传算法进行改进。

### 2 利用遗传算法改进 BP 人工神经网络

遗传算法(GA, Genetic Algorithm)<sup>[4,5]</sup>是一种非导数优化的优化方法,可以实现全局搜索,BP 学习采用的学习方法是梯度下降法,是基于导数的不稳定方法。因此,可以先用遗传算法对初始权值进行优化,在解空间中定位出较好的搜索空间。然后用 BP 算法在这些小的解空间中搜索出最优解。

### 3 算法描述与实现

根据以上思想,实现步骤如下:

1)建立单隐层 BP 人工神经网络,将前  $n$  年 GDP 和去年经济指标作为输入,期待输出明年的 GDP 指标。

2)利用遗传算法对 BP 神经网络的各权值进行优化,使误差最小。

3)用 hebb 学习方法,将历年数据输入 BP 神经网络进行学习。

4)以最后一组数据向后预测。

若选择某省工业景气指数,建筑业,股份有限公司,大型企业,国有控股企业,企业景气指数,利润指数为考虑因素,考虑三年的历史数据,则应该建立 10 输入,1 输出 BP 网络。根据 Konogorol 定理,隐层神经元设为  $2 * 10 + 1 = 21$  个。

归一化后作为基本指标。某省的 GDP 增长率基本指标见表 1(GDP 增幅已归一化,原始值为:11.22, 13.91, 10.71, 10.62, 9.22, 10.85)。

在 matlab 中用函数 newff 可方便地建立  $[10 - 21 - 1]$ BP 网络,利用遗传算法工具箱函数 gaoptimset 对网络做优化,利用返回的权优化 BP 网络,训练误差设为 0.001,最大次数设为 1000,用 train 函数进行逐年训练,可得到最终的神经网络。此网络已习得知识,带回前

表 1 某省六年的经济指标

年份	指标一	指标二	指标三	指标四	指标五	指标六	指标七	GDP 增幅
1	0.89	0.33	0.15	0.38	0.79	0.41	0.17	0.43
2	0.87	0.66	0.83	0.44	0.51	0.05	0.90	1.00
3	0.25	0.86	0.19	0.48	0.21	0.94	0.32	0.32
4	0.57	0.57	0.64	0.60	0.10	0.15	0.73	0.29
5	0.16	0.98	0.67	0.18	0.16	0.38	0.41	0
6	0.59	0.79	0.77	0.01	0.40	0.31	0.39	0.35

几年测试,误差平方和为 0.0038。这证明 GABP 神经网络的拟合特性是比较好的。

此网络预测第七年 GDP 增幅数据 0.5789,反归一化得 11.9320。

#### 4 规则抽取

以上结果表明人工神经网络的预测精度是比较高的,但是学得“知识”蕴涵在网络的连接权与结构中,如同一个“黑箱”,无法了解机器具体的分析过程。

问题是:需要从神经网络中获取一些关于机器习得知识的易于理解的、有意义的规则或者信息。原因很简单:人们对人工神经网络持有不信任感(虽然 Baum 和 Haussler<sup>[6]</sup>曾经指出,“如果一个神经网络可以为大量训练例产生正确的结果,那么可以相信它也能类似于训练例的未知示例产生正确结果”),这种不信任感来源于其“黑箱”特性。

1998 年,Tickle 等<sup>[7]</sup>指出,从神经网络中抽取规则是当前神经网络研究的一个重要课题。由此也引发了研究应用规则抽取技术的风潮,在各领域产生了一系列成果。

##### 4.1 保真度-精度窘境

另一个问题是,网络的存储知识极其复杂,人们究竟要获得什么程度的知识内容? FACC 框架<sup>[7]</sup>指出保真度、精度、一致性和可理解性是抽取工作需要追求的目标。然而文献<sup>[8]</sup>指出,对保真度的追求和对精度的追求在一定情况下是矛盾的。这种保真度和精度难以两全的现象称为保真度-精度窘境<sup>[9]</sup>(fidelity-accuracy dilemma)。在 GDP 分析中也正遇到这样的窘境。事实上,要追求满意的符号规则是比较困难的,符号规则抽取似乎更适用于模式分类<sup>[10]</sup>(如 Setiormo 的《Symbolic Representation Of Networks》)。而在实际决策中,人们往往关注的不是 IF-THEN,而是分析 GDP 数据时需考虑因素权重,以对关键因素优先决策,优先控制,以获得全局的最优化。

##### 4.2 求取关联系数

1993 年 Sestito 和 Dillon 提出了一个方法<sup>[11]</sup>用于分析人工神经网络习得知识。其初衷是构建符号规则,然而对于 GDP 这样的数据,想要离散化构造规则却是很困难的。以上已分析,最关心的是优先决策。根据要求,使用其计算各因子与 GDP 总量的相关系数。

首先将原神经网络的输出神经元作为附加输入神经元,然后利用扩展后的输入神经元和原输出神经元建立一个新的单隐层网络,并用 BP 算法对其进行训练。训练完成之后,对所有输入和附加输入神经元,根

据下式计算出它们之间的误差平方和 SSE,其中  $a$  为输入神经元, $b$  为附加输出神经元, $w_{aj}$  和  $w_{bj}$  分别为神经元  $a$  和  $b$  与隐层神经元  $j$  之间的连接权。 $SSE_{ab}$  度量了输入神经元  $a$  和输出神经元  $b$  之间的接近程度, $SSE_{ab}$  越小则说明输入  $a$  对输出  $b$  的作用越大。

$$SEE_{ab} = \sum_{j=0}^{\text{no of hidden units}} (W_{bj} - W_{aj})^2$$

据此 SSE 的值,可表征关联系数。

以上述某省份数据为例,在 matlab 中计算得相关系数为(已除去前三年的相关系数):

17.3428 16.2379 9.6463 18.9940 15.2624  
17.2695 14.2421

即表明了影响 GDP 的各因子与 GDP 关系程度,相关系数越大,此项对 GDP 影响越大。

由于这些因子与 GDP 为正相关,故可以参考按相关系数优先决策重要部分。

#### 5 结束语

从分析结果来看,此省的大型企业起主导,工业化程度高,企业景气指数对 GDP 的影响也比较大,因此决策者在调控 GDP 发展时需要注意发挥主要因素的主导作用,有效,快速地进行调控活动。

机器学习是一种第三方的分析方法,客观全面。文中结合规则抽取技术,介绍了人工神经网络建立、改进、分析方法,分析各影响因子的影响程度,无疑具有很强的现实意义。

#### 参考文献:

- [1] McCulloch W S, Pitts W H. A logical calculus of the ideas immanent in neuron activity [J]. Bulletin Mathematical Biophysics, 1943(5): 115-133.
- [2] Hechi Nielsen R. Theory of the back propagation neural network [J]. Int. J. Conf. on Neural Network, 1989(1): 593-605.
- [3] 张玲, 吴福朝, 张钺. 多层前馈神经网络的学习和综合算法 [J]. 软件学报, 1995, 6(7): 440-448.
- [4] Goldberg D E. Genetic Algorithm in Search, Optimization and Machine Learning [M]. New York: Addison Wesley, 1989.
- [5] 王小平, 曹立明. 遗传算法理论应用与软件实现 [M]. 西安: 西安交通大学出版社, 2002.
- [6] Baum E B, Haussler D. What size net gives valid generalization [J]. Neural Computation, 1989, 1(1): 151-160.
- [7] Tickle A B, Andrews R, Golea M, et al. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks [J]. IEEE Transactions on Neural Networks, 1998, 9(6): 1057-1068.

(下转第 233 页)

型和所设计的遗传算法的实际运行效果,以 19 个需求点为例进行求解计算。种群的规模均为 40,运行代数为 100,交叉概率和变异概率分别为 0.7、0.4。车辆的容量  $C$  为 2000。以服务中心为坐标原点,各需求点的坐标如表 1 所示:各点需求量随机分布,且分布规律满足模型要求。

表 1 需求点坐标

Node $i$	0	1	2	3	4	5	6	7	8	9
Longitude	0	27	36	61	8	51	74	10	93	44
Latitude	0	10	76	1	42	20	72	5	14	37
Node $i$	10	11	12	13	14	15	16	17	18	19
Longitude	60	8	96	78	48	36	76	87	6	15
Latitude	2	87	13	11	23	38	55	33	9	86

采用文中设计的算法进行数值实验,计算得到的 7 条调度方案如下:

- Route 1[0, 8,2]  
Total Weight: 1242.00; Total Distance: 123.24
- Route 2[0, 3,20,12]  
Total Weight: 1182.50; Total Distance: 146.72
- Route 3[0,11, 4,13]  
Total Weight: 1490.20; Total Distance: 146.72
- Route 4[0, 5,19,16]  
Total Weight: 1373.40; Total Distance: 136.81
- Route 5[0, 17 ,7 ]  
Total Weight: 1243.10; Total Distance: 136.81
- Route 6[0, 4, 9 ,18]  
Total Weight: 1497.20; Total Distance: 124.29
- Route 7[0, 10,15,6]  
Total Weight: 1472.80; Total Distance: 62.73

(上接第 226 页)

参考文献:

[1] 刘鉴澄. 基于 Asp. net 技术的动态统计图处理编程[J]. 韶关学院学报,2004,25(9):27-30.

[2] 乔平安. 仿 TeeChart 控件的统计图控件的设计与实现[J]. 西安邮电学院学报,2007,12(9):65-69.

[3] 刘院春,郑向宏. TeeChart 控件的结构分析与应用[J]. 计

(上接第 229 页)

[8] Andrews R, Diederich J, Tickle A B. Survey and critique of techniques for extracting rules from trained artificial neural networks[J]. Knowledge - Based Systems, 1995, 8(6):373-389.

[9] Zhou Z- H. Rule extraction:using neural networks or for neural networks[J]. Journal of Computer Science and Technolo-

4 结束语

通过对各需求点运输量  $D_i$  分布规律的深入探讨,分析了分布参数对随机需求车辆调度的影响,在建立数学模型的基础上设计了问题求解的并行遗传算法,对 19 个需求点的随机车辆调度问题进行了数值求解。求解结果表明:本节所提得并行遗传算法对车辆类型相同的需求随机的车辆调度问题能够取得满意的调度方案。但由于车辆调度问题还可以衍生出很多更为复杂也更贴近实际应用的问题<sup>[5,6]</sup>,如多个供应点、多车队、时间窗、多重运输需求、多种运输环节、多重交通网络、多重约束条件,这些都是带中转点联盟车辆调度问题必须综合考虑的问题,也是今后进一步研究的方向。

参考文献:

[1] Teodorovic D, Pavkovic G. A simulated annealing technique approach to the vehicle routing in the case of stochastic demand[J]. Transportation Planning and Tehnology,1992(16):261-273.

[2] 邹谷山. 车辆调度问题的遗传算法研究[D]. 广州:广东工业大学系统工程研究所,2005.

[3] 倪勤,袁健,刘晋. 随机需求的车辆路线问题的新模型[J]. 运筹与管理,2001,10(3):74-79.

[4] 封全喜,刘诚. 物流配送车辆路径问题的并行遗传算法研究[J]. 铁道工程与科学学报,2005,2(4):88-91.

[5] 蔡延光,钱积新,孙优贤. 多重车辆调度问题的模拟退火算法[J]. 系统工程理论与实践,1998(10):11-15.

[6] 蔡延光,钱积新,孙优贤. 多重车辆调度问题基于双表的并行表搜索算法[J]. 系统工程理论与实践,1998(11):20-26.

算机系统应用,2004,16(1):57-59.

[4] 朱玲,武玉强,张启宇. TeeChart 实现工控领域的实时曲线和历史曲线的方法[J]. 工业控制计算机,2005,18(8):49-50.

[5] 白鹏. Visual Basic 编程实例与技巧[M]. 北京:科学出版社,2003:142-161.

[10] 陈文伟,黄金才,赵新昱. 数据挖掘技术[M]. 北京:北京工业大学出版社,2002.

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