

基因表达式编程及其在混凝土徐变分析中应用

何鲜峰^{1,3,4}, 顾冲时^{1,2}, 谷艳昌^{1,2}, 陈睿^{1,2}

(1. 河海大学水电学院, 南京 210098; 2. 水资源高效利用与工程安全国家工程研究中心, 南京 210098; 3. 黄河水利科学研究院, 郑州 450003; 4. 水利部堤防安全与病害防治工程技术研究中心 郑州 450003)

摘要:混凝土徐变分析在混凝土结构应力计算和变形分析中有着重要意义。由于影响混凝土徐变的因素较多, 徐变函数呈现高度非线性, 传统回归方法难以取得令人满意结果。基因表达式程序设计方法, 吸取了遗传算法和基因编程两者的优点, 在解决非线性拟合问题方面表现出非凡的数据挖掘能力与优势。在基因表达式编程理论分析基础上, 利用该方法建立了混凝土徐变函数知识挖掘模型。实例表明, 该模型比常规方法得到的函数有更高的精度。

关键词:基因表达式编程(GEP); 混凝土徐变; 知识挖掘

中图分类号: TP311 文献标识码: A 文章编号: 1006-7329(2008)04-0082-05

The Application of Gene Expression Programming in Concrete Creep Analysis

HE Xian-feng^{1,3,4}, GU Chong-shi^{1,2}, GU Yan-chang^{1,2}, CHEN Rui^{1,2}

(1. College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing, 210098, P. R. China; 2. National Engineering Research Center of Water Resources Efficient Utilization and Engineering Safety, Nanjing, 210098, P. R. China; 3. Yellow River Institute of Hydraulic Research, Zhengzhou 450003, P. R. China; 4. Research Center on Levee Safety & Disaster Prevention Ministry of Water Resources, Zhengzhou 450003, P. R. China)

Abstract: The creep analysis of concrete is of great importance in stress calculation and deformation analysis of concrete structures. Since many factors affect concrete creep, the creep function is highly nonlinear, making it difficult to obtain satisfactory results using the traditional regression method. The gene expression programming (GEP) method, which combines the advantages of GA and gene programming, shows great promise and capability of data mining in nonlinear fitting. Based on theoretical analysis of GEP, the data mining model of the creep function was established by using GEP. Examples indicate this model is more accurate than functions obtained by conventional methods.

Key words: Gene Expression Programming (GEP); concrete creep; knowledge mining

混凝土徐变规律研究在混凝土结构应力、变形、预应力损失计算方面都有重要意义, 徐变函数的精确与否直接影响到上述计算结果的精度。由于徐变是混凝土的一种特殊属性, 对水泥品种、骨料品种、水灰比、龄期(养护龄期和加荷龄期)、荷载大小等多种因素比较敏感, 徐变函数 $C(t, \tau)$ 呈现高度非线性, 常规回归方法难以取得满意的效果。因此, 国内外学者在对大量混凝土徐变资料分析、归纳和总结的基础上提出了多种徐变函数表达形式^[1-4], 然而这些经验公式普遍存在

待定系数较多, 实际应用时系数不易确定的难题。虽然神经网络和演化算法分别在建立非线性关系和参数识别方面可以取得较好的效果^[5], 但神经网络在训练样本时容易陷入局部最优收敛, 且提供的是一个黑箱系统, 使用不便, 而演化算法受所选徐变函数样式的限制, 有时效果并不理想。而基因表达式编程(GEP, Gene Expression Programming)知识挖掘技术作为一种新兴的知识挖掘手段, 目前已在股票投资、信息系统设计、煤矿瓦斯监控和 SARS 疫情分析领域得到成功

* 收稿日期: 2008-03-20

基金项目: 国家自然科学基金(50579010, 50539010, 50539110); 973 基础研究发展计划(2007CB714103); 江苏省高校研究生科技创新计划(1044-B06015); 国家科技支撑计划(2006BAC14B03); 中国水电工程顾问集团公司科技项目(CHC-KJ-2007-02); 国家自然科学基金委、二滩水电开发有限责任公司雅砻江水电开发联合基金(50539030-1-3); 黄科院院长基金(HKY-JBYW-2008-13)

作者简介: 何鲜峰(1974-), 男, 工程师, 河海大学博士研究生, 主要从事水工建筑物安全监控研究与应用。(E-mail): hexf@hhu.edu.cn

应用^[6-10],可以比较完美的解决混凝土徐变这类高度非线性函数挖掘问题,并可在全域内达到最优。论文将就基因表达式编程及其在混凝土徐变分析中根据试验资料发现徐变规律的相关问题进行探讨。

1 GEP 程序设计概述

GEP 最初由 Candida Ferreira 于 2001 年提出^[11],此后得到迅速发展并逐步完善。GEP 来源于遗传算法(GA, Genetic Algorithms)和基因编程(GP, genetic programming),继承了二者的优点。在 GEP 进化过程中,既保持了 GA 算法遗传操作的便捷性,又拥有 GP 算法的非线性树结构,实现了利用简单编码解决复杂问题的目的^[12]。

1.1 GEP 的基因和染色体结构

基因表达式编程采用定长编码的符号串作为遗传基因编码,编码由头部(head)和尾部(tail)组成。头部由函数操作符(来自于函数集:运算符和初等函数)或终结符(来自于终结符集:自变量和常数)组成,而尾部仅含有终结符。头部长度 h 依据具体问题选定,尾部长度 l 则是 h 和 n 的函数,并由下式决定。

$$l = h \cdot (n - 1) + 1 \tag{1}$$

其中 n 为所需参数量最多的函数的参数个数。例如,开方和对数运算, n 取 1; 而加减等四则运算, n 取 2。GEP 中基因的这种结构避免了遗传操作中产生大量的无效编码,提高了算法的执行效率。

考虑由 $\{+, -, *, /, Q, S, C, a, b\}$ 构成的基因(其中 Q 表示开平方, S 和 C 分别表示正弦和余弦函数)。这里 $n=2$, 如果取 $h=10$, 则由式(1)可得 $l=11$, 基因总长度为 $10+11=21$ 。则式 $(\sin a - \cos b/a) + \sqrt{a} \cdot b$ 的基因编码可表示为(尾部用黑斜体表示):

$$012345678901234567890 + - * S/QbaCaabaabbbaaba \tag{2}$$

要将其变换成表达式树(Expression Tree, ET), 只要按照从左到右顺序读取基因, 并按照层次顺序构成表达式树。依照上述规则解析基因(2)得到的树如图 1 所示。该例中表达式长度为 12 个字符, 基因后部的非编码区域为程序的进化提供了很大空间。

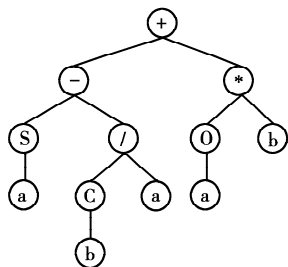


图 1 基因表达式树形结构

在 GEP 中,若干个类似的等长基因按照一定的组合方式构成了 GEP 染色体(Chromosome)。这种组合视具体情况而定,可以是逻辑运算,也可以是四则运算。程序运行时,基因的数目以及基因的头部长度都是事前选定的。染色体中每个基因片段可以解码成一个子表达式树(Sub-Expression Tree, Sub-ET),多个子表达式树构成了更复杂的多子树(Multi-subunit ET)表达式树^[11]。GEP 染色体的这种特殊结构以及丰富的遗传算子为 GEP 解决复杂问题提供了基本保证。

1.2 适应度函数和选择算子

在 GEP 的符号回归和函数发现过程中,所有的进化计算算法都需要对代表个体的问题解答进行评价。由于解答是一段程序,确切的说是一个表达式,对表达式进行评价,就是要评测利用该表达式计算得到的数据和目标数据的接近程度。Candida 提出了两种评价模型:

$$f_i = \sum_{j=1}^{C_i} (M - |C_{(i,j)} - T_{(j)}|) \tag{3}$$

$$f_i = \sum_{j=1}^{C_i} (M - \left| \frac{C_{(i,j)} - T_{(j)}}{T_j} \cdot 100 \right|) \tag{4}$$

其中, M 是适应度的取值范围,为一常数; $C_{(i,j)}$ 为由第 i 个染色体解析的公式计算第 j 组数据集得到的估计值; T_j 为第 j 组数据集的目标值; C_i 为训练样本总数。

GEP 中对选择算子并无特殊要求,研究结果表明^[13],选择不同的选择算子在个体优选能力上没有明显差别,有些方法可能对解决某个问题表现较好,但对其他问题则可能表现欠佳。然而,为了防止超级个体独霸种群,在解决比较复杂问题时最好采用锦标赛选择算子。

1.3 染色体的遗传算子

如果说 GEP 的染色体结构是其解决问题的物质基础,那么染色体在遗传过程中丰富多样的操作算子则是 GEP 解决复杂问题的重要保证。通常 GEP 包括复制、变异、插串和重组等几大类遗传算子。

1.3.1 复制算子(Replication) 根据适应度函数计算结果和轮盘赌的选择结果,被选中的染色体被忠实的复制到下一代。通常,适应度高的个体更容易被复制到下一代。在挑选复制基因的过程中,被选中的基因组根据轮盘赌的结果决定了被复制的次数,并根据种群中个体的数量进行相应次数的旋转,从而保证了种群的大小维持恒定。

1.3.2 变异算子(Mutation) 变异操作可以出现在染色体的任何部位。当对染色体的每一位编码遍历,对其进行随机测试,当测试值满足设定的变异概率

时,重新产生该位编码。为保证染色体结构的完整性,如果变异发生在头部(head),可以重新选择所有的符号,否则只能选择终结符。

1.3.3 插串算子(Transposition and insertion Sequence)

GEP 的转位因子(transposable elements)是能够被激活并跳跃到染色体其他地方的基因片段。针对转位因子,GEP 中有 3 种插串操作方式:1) S 插串(insertion sequence elements, IS elements)。该方法随机在基因中选择一段子串(起始位置、长度随机选择),然后将该子串插入到基因头部第一个位置之外的任何位置,将头部的其他符号向后顺延,并截掉超过头部长度的编码;2) RIS 插串(Root IS elements ,RIS)。RIS 插串是选择一个第一的位置是函数的子串,插入到染色体的起始位置,将头部的其他符号向后顺延,并截掉超过头部长度的编码;3)整个基因转移到染色体的起始位置,并称之为基因插串。

1.3.4 重组算子(Recombination) GEP 中有三种重组方式:单点重组(One-point recombination),两点重组(Two-point recombination)和基因重组(Gene recombination)。但基因重组并没有形成新的基因,只是两个父染色体的随机配对并交换对应部分的染色体编码。

1)单点重组 两条父染色体随机选择一个交换点,并彼此交换该点之后的所有代码,交换后的染色体作为子代进入下一轮进化过程;

2)两点重组 两条父染色体随机选择两个交换点,并彼此交换两点之间的所有代码,形成两个子代染色体;

3)基因重组 该操作只作用于多基因的染色体。随机选择一个基因,交换两个父代染色体的相对应的基因。

2 基于基因表达式程序设计的徐变规律知识挖掘模型

2.1 模型参数的确定

虽然影响混凝土徐变的因素较多,但对于使用材料、水灰比及养护条件一定的结构而言,决定徐变度的主要因素是持荷时间($t-\tau$)和加荷龄期 τ , t 为观察时刻。因此,与传统模型一样,GEP 建模选持荷时间和加荷龄期为徐变度函数的参数项。

2.2 GEP 算法流程

GEP 的算法流程与 GA 和 GP 有某些相似的地方,其主要结构流程如下:

Step1 创建初始种群,随机生成一定数量的染色

体,每个染色体含有若干个基因片段(基因个数一旦选定,运行过程不再更改),各个基因表达式之间通过连接函数(Link Function)连接。

Step2 根据选定的适应度函数和训练样本,计算每个个体的适应度。不同的适应度函数对种群进化过程的影响不同,经过比较,本文使用如下形式的适应度函数。

$$\alpha_i = \frac{1}{C_t} \sum_{j=1}^{C_t} (C_{(i,j)} - T_{(j)})^2 \quad (5)$$

$$f_i = 1\ 000 \frac{1}{1 + \alpha_i} \quad (6)$$

Step3 根据适应度计算值,判断是否满足收敛条件。若满足收敛准则,结束程序;否则执行下一步。

Step4 根据个体的适应度值和复制概率决定保留到下一代的个体。

Step5 根据选定的变异、插串、重组概率和选择算子,执行相应的遗传操作,生成新的种群。

Step6 返回第二步。

3 算例分析

为检验 GEP 算法的有效性,根据上述算理论编制相应的计算程序,对混凝土徐变规律进行了研究分析。

算例 1:对某大坝混凝土徐变试验资料进行知识挖掘处理,该大坝混凝土配合比见表 1。样本数据被分为两部分,其中的 126 组作为训练样本,另外不同加载龄期的 12 组数据作为检验样本。建模过程中染色体及遗传算子的选择概率等参数设置见表 2。计算结束后对适应度最优的染色体解码,得到徐变度计算公式为:

$$C(t, \tau) = \frac{8.404\ 694(10.422\ 058\tau - 8.583\ 832)}{\tau^2} - \frac{\{2.669\ 8\ \tau \ln[(t - \tau)\tau]\}}{(9.814\ 27t)} + \ln\{0.433\ 881\ 2[(t - \tau)^2 + \tau - 9.981\ 873]\} \quad (7)$$

表 1 试验样本混凝土配合比

水灰比	0.48	胶材用量/	水泥	182
用水量	96	(kg · m ⁻³)	硅粉	16
砂率/%	26	外加剂/	SW-1	0.8
含气量	4.1	%	SP169	0.02
混凝土配比(胶材:砂:石)			1:2.869:8.172	

表 2 建模参数设置

种群及染	种群大小	45	基因个数	3
色体结构	基因头长度	8	基因连接函数	加(+)
	复制	0.1	基因重组	0.1
遗传算	变异	0.05	IS 插串	0.1
子概率	单点重组	0.3	RIS 插串	0.1
	两点重组	0.1	基因插串	0.1

图 2 为 GEP 建模得到的模型(记为 MGEP)和根据文献[2]用回归方法求解相关参数得到的 7 参数模型(记为 M7)预测结果,表 3 为两种模型对检验样本的预测结果。分析图 2 中各加载龄期的预测结果可以发现, MGEP 模型明显优于常用的 M7 模型;表 3 中除一

组 MGEP 模型预测值稍逊于 M7 模型之外,其他预测结果 MGEP 模型均优于 M7 模型,检验样本预测结果的相对误差和相关系数进一步证实了 MGEP 模型的优越性。

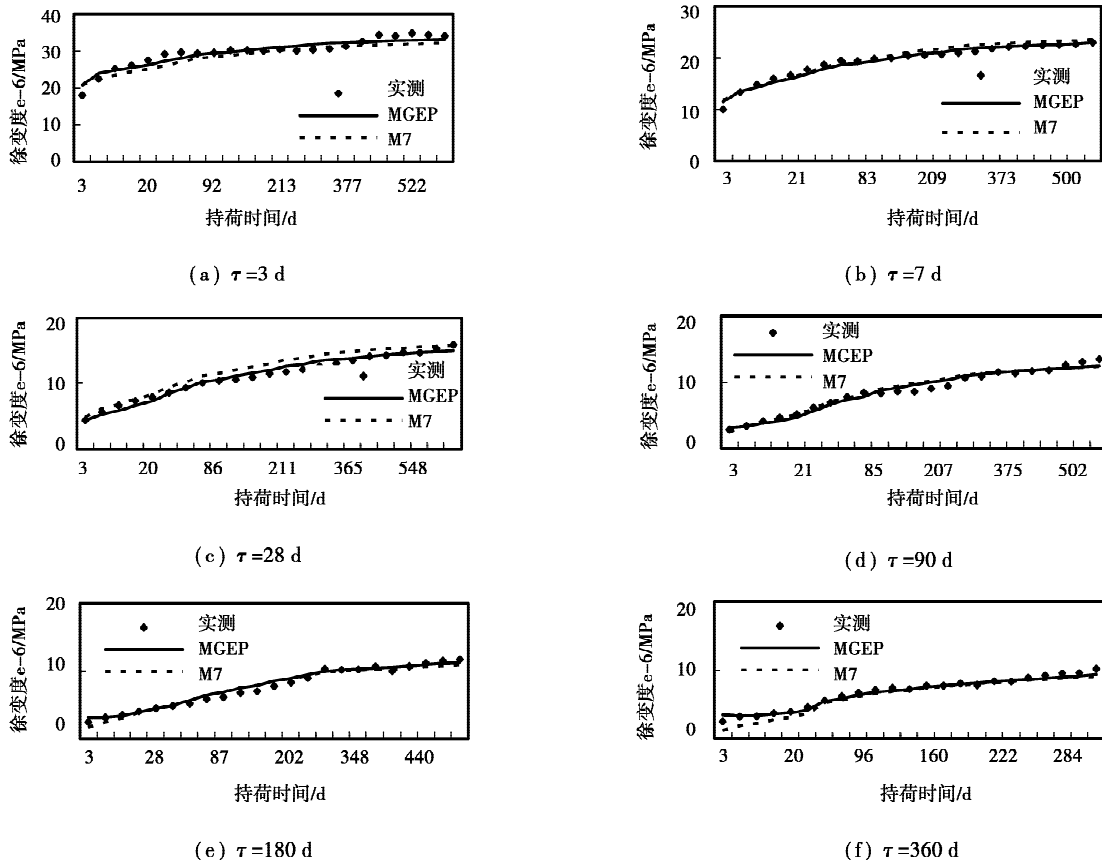


图 2 算例 1 徐变拟合曲线

表 3 检验样本 GEP 和 Model 模型预测结果对比表(10⁻⁶/MPa)

t- tau/d	tau/d	实测值	预测值		相对误差/%		相关系数	
			GEP	M7	GEP	M7	GEP	M7
458	3	34.31	32.59	31.66	5.02	7.71		
503	3	34.04	32.78	31.83	3.71	6.48		
209	7	20.54	20.83	21.47	-1.39	-4.52		
232	7	20.65	21.04	21.69	-1.89	-5.05		
211	28	11.76	12.63	13.61	-7.39	-15.73		
240	28	12.13	12.91	13.90	-6.45	-14.56	0.998	0.992
375	90	11.51	11.43	11.51	0.65	0.00		
407	90	11.34	11.63	11.67	-2.56	-2.95		
52	180	5.20	5.68	5.72	-9.29	-9.96		
72	180	5.85	6.39	6.43	-9.32	-9.91		
127	360	7.36	6.96	6.80	5.44	7.57		
302	360	9.58	9.12	8.93	4.83	6.76		

算例 2:对来自文献[3]的 515 工程混凝土的徐变试验资料进行 GEP 建模分析(该模型染色体基因个数为 4,其他参数同表 2),得到 GEP 模型徐变度公式为:

$$C(t, \tau) = \ln\left[(t - \tau)^2 + \frac{\tau}{(t - \tau)}\right] - \frac{\sqrt{\tau}}{2.915\ 039} + 14.199\ 765\ 6 + 2[\ln(t - \tau) - \ln \tau] + \frac{\{8.505\ 951\ln[3(t - \tau) - 4.803\ 711]\}}{\sqrt{\tau}} - \frac{9.991\ 913}{(\sqrt{t - \tau} - \tau - 8.165\ 374)} \quad (8)$$

图 3 为 GEP 模型和文献[3]的 7 参数模型对比结果,从中可以看到,GEP 模型可以得到更好的计算效果。另外,就计算值和试验观察值的相关性而言, R_{MGEP}=0.995 8,而 R_{M7}=0.984 4,这也进一步说明了 GEP 模型在混凝土徐变规律挖掘中可以达到相对更高的计算精度。

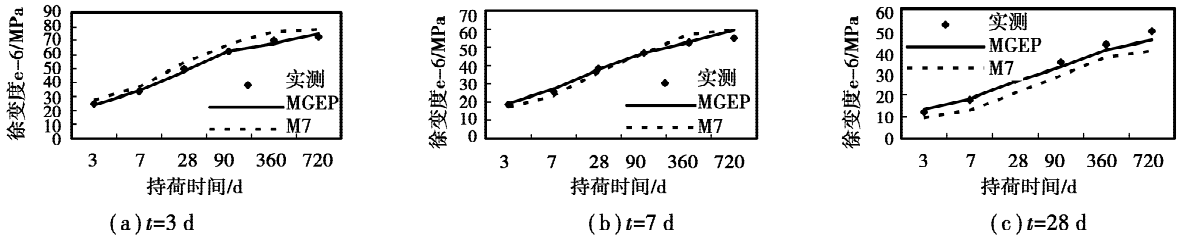


图 3 515 工程混凝土徐变拟合曲线

4 结 语

GEP 方法进行知识发现,不需要知道给定条件和待求解目标之间的内在关系,只要实验数据或训练样本足够丰富,就可以比较准确的发现求解条件和追求目标之间的函数关系。实例计算结果表明,GEP 方法在根据实验数据挖掘混凝土徐变度函数这样高度非线性函数方面,与传统方法相比,表现出明显的优势。可以预见 GEP 方法也将会在其他工程和非工程领域的知识挖掘、模式识别等方面发挥更大的作用。

参考文献:

[1] 康崇钊. 混凝土的徐变力学与实验研究[M]. 北京:水利电力出版社,1982.

[2] 朱伯芳. 混凝土的弹性参数、徐变度与应力松弛[J]. 水利学报,1985(9):54-61.
ZHU Bo-fang. The elastic parameters, creep and stress relaxation of concrete[J]. Journal of Hydraulic Engineering,1985(9):54-61.

[3] 惠荣炎,黄国兴,易冰若. 混凝土的徐变[M]. 北京:中国铁道出版社,1988:159-192.

[4] 陈志华. 混凝土徐变预测模型的对比分析[J]. 桥梁建设,2006(5):76-78,81.
CHEN Zhi-hua. Comparative analysis of concrete creep prediction models[J]. Bridge Construction, 2006(5):76-78,81.

[5] 陈志华,陕亮,关富玲. 基于演化程序的混凝土徐变参数识别[J]. 长江科学院院报,2005,4(22):47-49.
CHEN Zhi-hua, SHAN Liang, GUAN Fu-ling. Identification of concrete creep parameters based on evolution programs[J]. Journal of Yangtze River Scientific Research Institute,2005,4(22):47-49.

[6] 黄隆胜,肖士斌. 基因表达式编程在 SARS 疫情分析及预测中的应用[J]. 计算机工程,2007(4):45-48.
HUANG Long-sheng, XIAO Shi-bin. Application of SARS epidemic situation analysis and forecast based on gene expression programming[J]. Computer Engineering, 2007(4):45-48.

[7] 李曲,蔡之华,朱利,等. 基因表达式程序设计方法在采

煤工作面瓦斯涌出量预测中的应用[J]. 应用基础与工程科学学报,2003,12(1):50-54.

LI Qu, CAI Zhi-hua, ZHU Li, et al. Application of gene expression programming in predicting the amount of gas emitted from coal face[J]. Journal of Basic Science and Engineering, 2003,12(1):50-54.

[8] 余弦,唐常杰,张欢,等. 基于基因表达式编程挖掘中医方证关系[J]. 计算机应用,2005,11(25):2 679-2 680.
YU Xian , TANG Chang-jie, ZHANG Huan, et al. Mining formula-syndrome relation shi Pin traditional Chinese medicine with gene expression programming[J]. Computer applications, 2005,11(25):2 679-2 680.

[9] 牛红惠,何芳. 基于基因表达式编程的数据挖掘技术在股票中的应用[J]. 福建电脑,2006(10):109-110.
NIU Hong-hui, HE Fang. Programming based on the gene expression data mining technology in the stock of research and application[J]. Fujian Computer,2006(10):109-110.

[10] 陆昕为,蔡之华,陈昌敏,等. 基因表达式程序设计在信息系统建模预测中的应用[J]. 微计算机信息,2005(35):185-186.
LU Xin-wei, CAI Zhi-hua, CHEN Chang-min, et al. Ansheng application of gene expression programming in evolutionary modeling[J]. Control & Automation, 2005(35):185-186.

[11] CANDIDA FERREIRA. Gene expression programming: a new adaptive algorithm for solving problems [J]. Complex Systems, Vol. 13, issue 2: 87-129, 2001:1-22.

[12] 唐常杰,张天庆,左劼,等. 基于基因表达式编程的知识发现——沿革、成果和发展方向[J]. 计算机应用,2004,10(24):7-10.
TANG Chang-jie, ZHANG Tian-qing, ZUO Jie, et al. Knowledge discovery based on gene expression programming: history, achievements and future directions [J]. Computer Applications,2004,10(24):7-10.

[13] M MITCHELL, PT HRABER and J PCRUTCHFIELD. Revisiting the edge of chaos: evolving cellular automata to perform computations [J]. Complex Systems,1993,7:89-130.

(编辑 王秀玲)