文章编号:1000-5862(2014)03-0286-04

数据插值技术对基于遗传编程算法符号回归的影响研究

于　鹏[1],王　京[2]
(1. 渤海大学工学院,辽宁 锦州　121000;2. 国网锦州供电公司,辽宁 锦州　121000)

摘要:提出一种先采用样条函数插值对数据进行平滑处理,再采用基于遗传编程的符号回归方法进行数据分析的新方法. 与实验补充数据的方法相比,该方法实验成本低、风险小、受环境影响小,在电力电子技术领域具有广泛的适用性和较好的应用潜力.

关键词:符号回归;样条插值;遗传编程;数据拟合;电力电子
中图分类号:TP 311　　**文献标志码**:A

0　引言

回归分析是应用数学的方法,对大量的观测数据进行处理,得出比较符合事物内部规律的数学表达式的方法. 符号回归是一种建立在进化计算基础上的方法,主要用于数学表达式空间中进行误差最小化检索. 传统方法一般用于匹配确定函数形式的参数,符号回归既检索函数形式也检索参数[1]. 采用遗传规划方法的优点在于不需要给出具体的函数形式即可获得拟合的函数表达式,并且在初始群体足够大而且交叉和变异概率设置合理的情况下,不会陷入局部最优[2].

遗传规划方法需要将数据分为 2 个数据集,一部分数据集用于对数据进行训练,生成函数集合;另一部分数据集用于对训练结果误差进行判断. 在相对复杂情况下若数据量过少,会出现难以识别出数据之间对应关系的情况. 电力电子实验环境配置复杂,一些实验具有危险性,人力实验效率低下. 存在着实际实验难以操作或存在因为实验仪器设备的量程或分辨率限制造成数据空缺的情况. 若能够减少数据对实验装置的要求,则对降低电力电子装置开发成本、提高开发效率将具有现实意义. 若进行插值补充数据不影响符号回归结果,或者对结果影响在可以辨识范围内,则对于实验设计、实验数据的采集与实验数据补充均可以进行有针对性的缩减,以达到减少实验次数,节约人力物力的目的.

1　基本原理

采用数据插值的方法辅助数据分析方法,生成减少数据点的相对不完备数据集,对不完备数据序列进行 3 次样条平滑插值,扩展不完备样本数据量. 对平滑处理之前与之后的数据,分别进行基于遗传编程的符号回归分析辨识. 根据符号回归获得的表达式,分析样条函数平滑对于原始数据分析结果的影响.

插值基本原理是在离散数据的基础上补插连续函数,使得该连续曲线通过全部给定的离散数据点[3]. 在多项式拟合过程中,对每组相邻的数据点,用多项式去拟合数据点之间的曲线[4]. 插值的方法有临近点插值、线性插值、立方插值相和 3 次样条插值等,其中 3 次样条插值占用计算机计算资源较多,插值之后的曲线光滑程度较好. 符号回归也称为函数建模,根据给定的一组自变量和一组函数值,找出拟合函数关系式[5]. 这类似于参数回归,不同之处在于参数回归要事先给定具体函数形式,而符号回归则不需事先给定具体函数形式.

遗传编程基本过程是随机产生一个适合于给定环境的初始群体. 每个群体的个体都有一个适应度值,用遗传算法处理得到高适应度的个体,产生下一代的群体. 通过个体变异、择优实现代际进化,直到出现给定问题的解或近似解为止. 将遗传编程应用于符号回归过程中,可实现对数据表达式的自动识

收稿日期:2013-10-12
基金项目:辽宁省科学事业公益研究基金(2011004001)资助项目.
作者简介:于　鹏(1977-),男,辽宁锦州人,讲师,博士,主要从事电力电子方面的研究.

别. 遗传编程是一种相当有效的符号回归方法,它的优势源于其结构的灵活性,由于采用树形结构,可以描述层次化的问题,克服了传统方法中确定函数结构难的缺陷[6-15].

2　仿真实验和结果

为了验证数据插值对于基于遗传编程的符号回归方法在函数表达式自动识别过程中的影响,本文采用 Matlab 和 Eureqa 针对具体实例进行实验.

实验步骤: 通过 Matlab 生成数据,形成标准正弦信号数据集、延长时间步长的数据集以及针对延长时间步长数据集进行 3 次样条插值之后的数据集. 为了模拟较低分辨率的情况,通过增加时间步长来得到降低时间分辨率的数据集. 对于降低分辨率的数据集,采用 spline 函数进行 3 次样条插值,得到步长为 0.000 1 s 的插值之后的数据集. 将标准正弦信号数据集、延长时间步长的数据集、3 次样条插值之后的数据集导入到 Eureqa 软件中,进行识别. 比较分析辨识标准正弦信号数据集、延长时间步长的数据集、3 次样条插值之后的数据集得到的结果,分析 3 次样条插值对最后辨识结果的影响.

第 1 组实验采用标准正弦波进行表达式识别. 应用标准正弦波形表达式 $y = \sin(200\pi x)$ 采用 0.001 s步长生成数据集,部分数据见表 1.

表 1　标准正弦数据示例

序号	t/s	u/V
1	0	0
2	0.000 1	0.050 802 904 623
3	0.000 2	0.100 126 845 330
4	0.000 3	0.147 985 678 526
5	0.000 4	0.194 393 260 619
6	0.000 5	0.239 363 448 015
7	0.000 6	0.282 910 097 120
8	0.000 7	0.325 047 064 341
9	0.000 8	0.365 788 206 084
10	0.000 9	0.405 147 378 756

对数据集进行辨识,得到图像如图 1 所示. 图 1 中拟合曲线与数据点重合,点为数据校验点,曲线为拟合公式生成曲线. 经过 13 507 代遗传选择,识别结果表达式为 $y = 0.983\ 176\ 836\ 190\ 568 \cdot \sin(314.159\ 259\ 795\ 561x)$. 识别结果的绝对平均误差(MAE)为 0.008 432 059 4. 可见符号回归算法直接辨识得到标准正弦函数表达式.

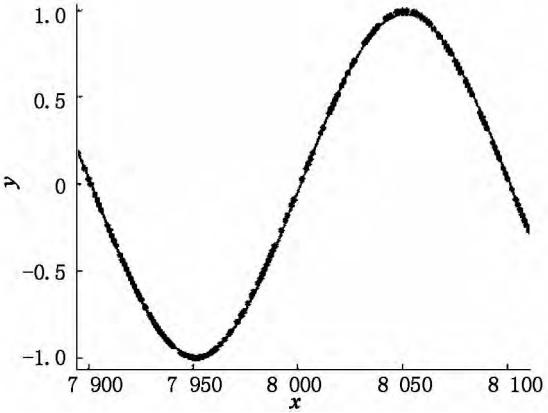

图 1　标准正弦曲线与识别结果

第 2 组实验应用标准正弦信号表达式 $y = \sin(100\pi x)$,为了模拟数据不完备的情况,采用 0.005 s步长生成数据集. 步长加大之后用获得的数据集进行辨识,结果如图 2 所示. 由图 2 可见,由于步长较大,三角函数如果用直线连接会形成折线结构. 在遗传编程代数为 36 622 代时,结果表达式为: $y = \sin(2\ 827.433\ 388\ 134\ 17x)$. 此时软件判断识别误差结果为 0,遗传编程迭代停止. 结果符号回归表达式函数形式正确,但是表达式辨识参数达到生成原始数据函数相应参数的 9 倍,结果表明数据集不完备造成的信息缺失影响了符号回归算法对结果表达式的识别,在软件识别函数表达式结果误差为 0,遗传编程计算停止情况下得到了参数误差较大的表达式.

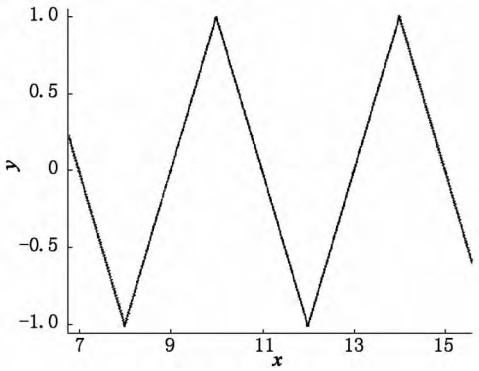

图 2　步长为 0.005 s 的正弦图像

第 3 组实验对以上数据集进行 3 次样条插值,步长为 0.000 1 s. 得到新数据集,其数据量增加为处理之前数据的 50 倍. 对此数据集进行辨识,程序运行3 590 245代之后得到表达式 $y = 0.983\ 611\ 337\ 2\sin(314.181\ 018\ 4x)$,识别图像如图 3 所示. 拟合曲线与数据点基本重合.

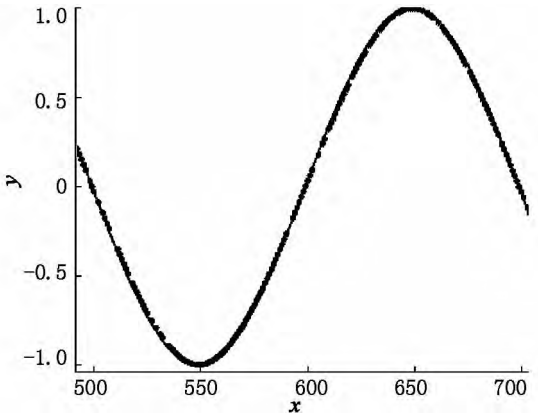

图 3　3 次样条插值之后的数据波形

在此基础上继续运行 $1.869\ 4\times10^{7}$ 代之后,得到了一组表达式:

$$y=0.106\ 863\ 521\ 853\ 543\cdot\cos(1.574\ 467\ 290\ 074\ 5+942.402\ 606\ 748\ 351x)+2.420\ 510\ 740\ 815\ 83\sin(\sin(0.000\ 360\ 502\ 039\ 173\ 734+314.151\ 924\ 420\ 076x))-1.145\ 495\ 122\ 478\ 77\cdot\sin(0.000\ 360\ 502\ 039\ 173\ 734+314.151\ 924\ 420\ 076x)\cdot\cos(-0.050\ 632\ 937\ 063\ 890\ 2\sin(942.402\ 606\ 748\ 351x)).$$

由识别表达式可见符号回归算法识别并未将样条函数多项式识别到最终正弦函数的表达式中. 插值对于符号回归识别的影响一方面表现在遗传代数增加,另一方面以上表达式中在标准表达式基础上引入了干扰项. 干扰项包括正弦叠加项:

$$2.420\ 510\ 740\ 815\ 83\cdot\sin(\sin(0.000\ 360\ 502\ 039\ 173\ 734+314.151\ 924\ 420\ 076x));$$

余弦叠加项:

$$0.106\ 863\ 521\ 853\ 543\cos(1.574\ 467\ 290\ 074\ 5+942.402\ 606\ 748\ 351x),\cos(-0.050\ 632\ 937\ 063\ 890\ 2\cdot\sin(942.402\ 606\ 748\ 351x)).$$

如果忽略三角函数嵌套项与系数较小项,会得到关系式 $y=1.145\sin(314.151\ 924\ 420\ 076x)$. 可见对干扰项进行处理之后的结果接近于生成数据的表达式.

第 4 组实验采用正弦信号叠加偶次谐波,应用表达式 $y=\sin(100\pi x)+\sin(200\pi x)$ 以步长 0.000 1 s数据生成数据集. 对数据集进行辨识,遗传代数 35 485 时识别表达式为

$$y=2\sin(314.159\ 265\ 700\ 214x)\cdot\cos(314.159\ 265\ 700\ 214x)+\sin(314.159\ 265\ 700\ 214x).$$

实验结果表明,符号回归程序辨识出了 2 次谐波表达式,表达式与参数都得到了拟合.

第 5 组实验采用正弦信号叠加偶次谐波表达式 $y=\sin(100\pi x)+\sin(200\pi x)$,采用步长 0.04 s 生成数据集. 对生成数据集进行识别,当遗传代数为 51 576代时,识别表达式 $y=1.579\ 500\ 620\ 611\ 4\cdot\sin(3.059\ 254\ 779\ 456\ 46\sin(628.318\ 529\ 789\ 666x))$.

得到以上结果之后,软件最终误差为 0,识别终止,拟合图像见图 4. 结果表明在数据量较少、数据未经过平滑处理的情况下,软件未能正确辨识数据表达式.

第 6 组实验采用正弦信号叠加偶次谐波,$y=\sin(100\pi x)+\sin(200\pi x)$,采用步长 0.04 s 生成数据集,对此数据集进行 3 次样条函数插值处理. 插值步长为 0.000 1 s,生成数据见表 2.

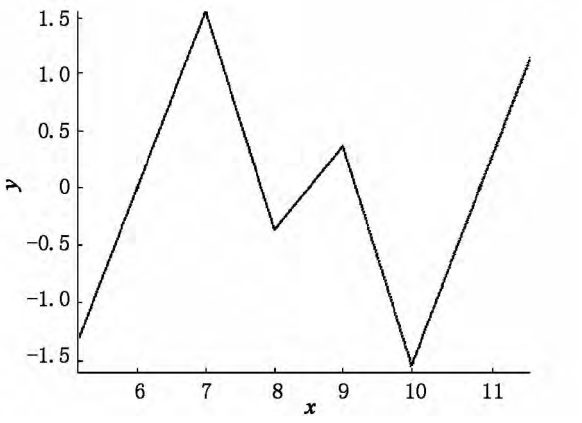

图 4 0.005 s 步长标准正弦波形叠加 2 次谐波数据

表 2 对标准正弦信号抽取数据插值之后的数据

序号	t/s	u/V
1	0.020 0	0
2	0.020 1	0.030 103 254 951
3	0.020 2	0.060 175 009 890
4	0.020 3	0.090 191 459 958
5	0.020 4	0.120 128 800 295
6	0.020 5	0.149 963 226 045
7	0.020 6	0.179 670 932 346
8	0.020 7	0.209 228 114 342
9	0.020 8	0.238 610 967 172
10	0.020 9	0.267 795 685 978

对处理之后经过插值的数据集进行识别. 识别的曲线见图 5. 遗传 6 573 449 代之后结果为

$$y=0.825\ 169\ 079\ 692\ 342\sin(0.009\ 433\ 089\ 463\ 483\ 58+628.117\ 725\ 210\ 552x)+1.153\ 528\ 693\ 910\ 25\cdot\sin(314.162\ 285\ 368\ 006x-0.257\ 101\ 196\ 775\ 551\cdot\sin(628.160\ 818\ 666\ 636x+0.179\ 099\ 070\ 127\ 807\cdot\sin(628.117\ 725\ 210\ 552x+6.461\ 067\ 953\ 423\ 04\times10^{-5}/x^{2}-0.636\ 350\ 741\ 395\ 767\sin(0.009\ 433\ 089\ 463\ 483\ 58+628.117\ 725\ 210\ 552x)))).$$

绝对平均误差为 0.003 8. 对识别结果,忽略参数较小项与表达式嵌套项,得到表达式

$$y=0.825\ 169\ 079\ 692\ 342\sin(628.117\ 725\ 210x)+1.153\ 528\ 693\ 910\ 25\sin(314.162\ 285\ 368x).$$

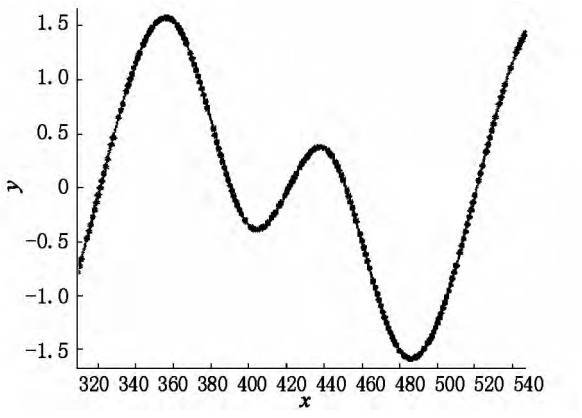

图 5 样条插值之后曲线

对标准正弦表达式进行比较识别的结果对于 2 次谐波叠加情况适用,对于插值之后符号回归辨识的表达式只要将参数较小项和表达式嵌套项忽略就可以得到精度较高的结果.

由以上实验可知样条插值对识别的影响在于: (i) 由于数据增加,识别时间延长; (ii) 由于样条插值过程,识别中在原函数基础上会增加干扰项. 干扰项一般表现为参数较小项与表达式嵌套项,其中嵌套项的产生,与符号回归表达式的生成规则有关.

对基本表达式的判断可以遵循以下原则: (i) 对主函数的识别,应忽略结果中系数较小的函数项; (ii) 应忽略多重嵌套项. 按以上原则处理之后的结果可以作为符号回归表达式.

3　结论

实验结果表明,采用数据插值技术进行数据处理辅助符号回归分析的方法是可行的. 采用该方法,可以在不进行实验补充数据的前提下对不完备数据集进行补充,应用补充后得到的新数据集进行符号回归辨识. 但是样条平滑也会对符号回归结果造成干扰,增加迭代次数. 样条平滑对符号回归的结果造成的干扰: (i) 参数识别的干扰; (ii) 函数形式的干扰,其中对正弦函数参数的干扰表现在幅度参数较实际值增加误差,对函数形式的干扰表现在函数主要成分的基础上增加函数多重嵌套结构. 虽然具有干扰,在特定情况下,如果实验数据无法补充或者补充实验数据成本较高,采用数据插值技术补充数据,通过增加计算机辨识时间来完成函数关系辨识不失为一种廉价高效、功能实用的数据处理手段. 在电力电子领域应用基于数据插值技术的数据预处理方法结合基于遗传编程的符号回归方法进行数据分析具有良好的前景.

4　参考文献

[1] Schmidt M, Lipson H. Distilling free form natural laws from experimental data [J]. Science, 2009, 324: 81-85.

[2] 夏炎. 遗传规划理论及其在符号回归中的应用 [D]. 上海: 上海交通大学, 2007.

[3] Schmidt M, Lipson H. Age fitness pareto optimization genetic programming theory and practice VIII [M]. Boston: Kluwer Academic Publishers, 2011.

[4] 封建湖, 聂玉峰, 王振海. 数值分析导教导学导考 [M]. 西安: 西北工业大学出版社, 2006.

[5] 陈杰. MATLAB 宝典 [M]. 北京: 电子工业出版社, 2010.

[6] 谢大同, 康立山. 符号回归的一种新算法 [J]. 系统仿真学报, 2007, 19(8): 1668-1671.

[7] Koza J K. Genetic programming [M]. Boston: MIT Press, 1992.

[8] 陈志卫. 遗传规划研究的现状及发展 [M]. 浙江工业大学学报, 2003(4): 153-159.

[9] 夏炎. 基于遗传规划的符号回归研究 [J]. 中国计量学院学报, 2006(6): 128-131.

[10] 王战权. 改进的遗传规划研究 [J]. 系统工程理论与实践, 2000(5): 66-69.

[11] Poli R. Parallel distributed genetic programming [M]. New Ynk: McGraw-Hill Press, 1999.

[12] 王小平. 遗传程序设计及其在符号回归问题中的应用 [J]. 同济大学学报: 自然科学版, 2001(10): 1200-1204.

[13] 金金宝. 遗传编程在符号回归中的应用 [J]. 计算机与数字工程, 2009(5): 13-16.

[14] Schmidt M, Lipson H. Coevolution of fitness predictors [EB/OL]. [2013-09-13]. http://creativemachines.cornell.edu/papers/TEC08_Schmidt.pdf.

[15] Blackledge J. Cryptography using evolutionary computing [J]. Signals and Systems Conference, 2013(6): 1-8.

The Research of the Influence of Data Interpolating Technology on Symbolic Regression Based on Genetic Programming

YU Peng[1], WANG Jing[2]

(1. College of Engineering, Bohai University, Jinzhou Liaoning 121000, China;
2. Jinzhou Electric Power Supply Company of State Grid Corp China, Jinzhou Liaoning 121000, China)

Abstract: A new method which use data interpolation technology to preprocessing data has been proposed. After data preprocessing, the data is processed by the symbol regression method. Comparing to traditional method that get data by practical experiment, this method is cheap and low-risk. The method is independent from environment. In the field of power electronics, this method has strong applicability and a good application prospect.

Key words: symbolic regression; data interpolating; genetic programming; data fitting; power electronics

(责任编辑: 冉小晓)