基于粒子群优化算法的测光红移回归预测

穆永欢[1]，邱 波[1*]，魏诗雅[1]，宋 涛[1]，郑子鹏[1]，郭 平[2*]

1. 河北工业大学电子信息工程学院，天津 300400
2. 北京师范大学系统科学学院，北京 100875

摘 要 星系的红移在天文研究中极其重要，星系测光红移的预测对研究宇宙大尺度结构及演变有着重要的研究意义。利用斯隆巡天项目发布的SDSS DR13的150 000个星系的测光及光谱数据进行分析，首先根据颜色特征并基于聚类的方法对星系进行分类，由分类结果可知早型星系的占比较大。对比了三种不同的机器学习算法对早型星系进行测光红移回归预测实验，并找出最优的方法。实验中将星系样本中u，g，r，i，z五个波段的测光值以及两两做差得到的10个颜色特征作为输入数据，首先构建BP网络，使用BP算法对星系的测光红移进行回归预测；然后利用遗传算法(GA)优化BP网络各层参数，将优化后的GA-BP算法应用于早型星系的回归预测试验中。考虑到GA算法的复杂操作会影响预测效率，并且粒子群算法(PSO)不仅稳定性高且操作简单，因此将粒子群算法应用到星系样本中早型星系的测光红移回归预测实验中，进而采用粒子群算法优化BP网络(PSO-BP)。实验中将光谱红移作为期望值，采用均方差(MSE)作为误差分析指标来评判三种算法的精度，将PSO-BP回归预测结果与BP网络模型、GA-BP网络模型进行比较。由实验结果可知，BP网络的MSE值为0.001 92，GA-BP网络的MSE值0.001 728，PSO-BP网络的MSE值为0.001 708。实验结果表明，所用到的PSO-BP优化模型在精度上优于BP神经网络模型和GA-BP神经网络模型，分别提高了11.1%和1.2%；在效率上优于传统的K近邻(KNN)测光红移估计算法，克服了KNN算法中遍历所有数据样本进行训练的缺点并且其泛化性能优于其它BP网络优化模型。

关键词 测光红移；粒子群优化；粒子群算法优化BP网络；BP神经网络；GA-BP神经网络

中图分类号：P157.2 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)09-2693-05

引 言

星系可以被视作为一个由多种物质组成的复杂天体运行系统，目前人类利用先进的巡天望远镜已经可以在宇宙中观测到的星系总数超过1 000亿个。所以星系在天文研究中极其重要。

星系的红移是指在光源与观测点之间的距离发生变化时，光波在波长上的变化为距离越远波长越长，在频率上则表现为距离越远频率越低。在可见光波段，光谱的波长由于距离变远将表现为谱线朝红端移动，这种现象称为“红移”。星系的红移可分为两类：①光谱红移定义为$z=\frac{\lambda-\lambda_0}{\lambda_0}$，($\lambda$表示星系的同一谱线波长，$\lambda_0$表示地球上光源中某一谱线波长)其测量主要是由光纤光谱仪完成。②测光红移是指通过大型的CCD相机对星系在多个波段进行曝光和数据采集处理，然后确定得到测光红移。测光红移的估测主要有：①光谱能量分布SED[1](spectral energy distribution，SED)拟合法：首先需要建立一系列模板，供实际观测得到的星系进行颜色比对以确定星系的测光红移；②训练集方法：将训练数据作为模型训练的输入特征，而光谱红移因为高精度的特点，则作为目标值来约束模型，使用机器学习的方法，最终通过训练拟合两者之间的关系[2]；③事例学习法：与训练集方法相似，不同之处在于事例学习法不需要光谱红移作为目标值。

近年来针对星系测光数据呈爆发式增长，为同源天体提供了多波段数据来源[3]。SDSS最新发布的SDSS DR13(斯隆数字巡天DR13数据库)中包含两亿多个星系的测光数据[4]。测光数据的急剧增长使应用高效和准确的机器学习算法进行测光估计成为必然趋势。目前，应用最广泛的训练集方法包

收稿日期：2018-07-19，**修订日期**：2018-11-25

基金项目：国家自然科学基金委员会-中国科学院天文联合基金项目(U1531242)，河北省科技支撑计划项目(15212105D)资助

作者简介：穆永欢，1993年生，河北工业大学硕士研究生 e-mail：1944987108@qq.com

*通讯联系人 e-mail：qiubo@hebut.edu.cn；pguo@bnu.edu.cn

括最近邻[5]、随机森林[6]、支持向量机方法[7]等。

本文利用机器学习算法中的 BP 神经网络对星系的测光红移进行回归预测，进而使用粒子群算法(particle swarm optimization，PSO)对 BP 神经网络进行优化(简称 PSO-BP)，将优化后的 PSO-BP 网络用于对早型星系进行测光红移的回归预测中。最后将 PSO-BP 的实验结果分别与 BP 神经网络和基于遗传算法(genetic algorithm，GA)的 BP 神经网络(简称 GA-BP)的实验结果进行分析比较。

1 数据分析

SDSS 是至今规模最大，数据量最丰富的数字巡天项目。在 SDSS 巡天数据中，具有红移的样本包括：主星系样本(main galaxy sample，MGS)，亮星系样本(luminous red galaxy，LRG)和类星体样本。最新发布的 SDSS-DR13 中，光谱数据已经达到四百多万个，其中包含 240 多万个星系的光谱数据。本文从 SDSS-DR13 中选取 u，g，r，i，z 五个波段的测光数据和两两相减得到的颜色特征作为输入特征，将光谱红移量作为期望值，最后比较实验误差来评估回归模型精度。

星系分类有很多标准，包括传统的形态学分类、光谱分类和颜色分类等[8]。由于星系的颜色与光谱能量分布密切相关，故本文实验根据颜色特征，基于聚类的方法对星系进行分类，可为早型星系和晚型星系。根据颜色进行聚类，其 u-r 的分布直方图如图 1 所示。

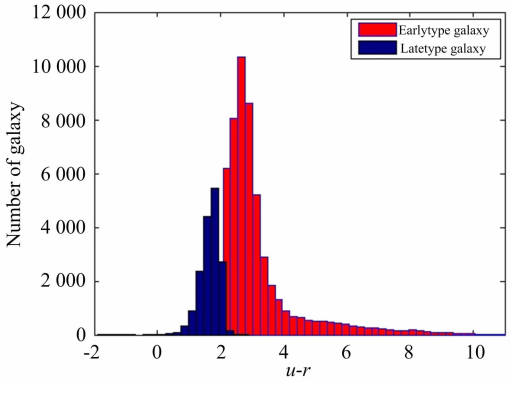

图 1 星系聚类直方图

Fig. 1 Histogram of galaxy clustering

由星系聚类的分布直方图可以看出，在星系中，早型星系占有较大的比例，故为了减少样本复杂性，仅将早型星系作为输入样本进行分析。

2 三种神经网络算法

2.1 BP 神经网络

BP 神经网络正是一种利用误差的逆向传导来自动化调整模型参数进行网络训练的多层前馈式网络[9]，输入信号经输入层输入，通过隐含层内部计算得到输出信号值与期望输出值的误差，再将误差信号反向沿着输出层→隐含层→输入层进行传播。在传播时，通过误差梯度下降方法，更新神经网络各层神经元的权值和阈值[10]。在信号和误差通过正、反向传播中不断更新各层权值，直至 BP 神经网络的误差输出下降到目标误差值，则停止训练。

BP 神经网络操作简单，然而算法的稳定性比较差，收敛时间长，可以采用相应的优化算法对 BP 神经网络进行优化，从而提高算法的稳定性。

2.2 GA-BP 神经网络算法

遗传算法(GA)是一个高度并行的自适应检测法。遗传算法能够在数据空间进行全局寻优，且高度收敛，故可对 BP 神经网络进行优化，提升效率[11]。

基于遗传算法的 BP 神经网络(GA-BP)优化步骤如下：

(1)种群参数初始化。

(2)求解适应度函数。

(3)进行交叉运算、变异运算，找到最优种群。

(4)计算新种群的适应度值，当满足条件时停止优化。否则返回步骤(3)继续优化。

通过 GA 优化的 BP 神经网络，可避免 BP 算法陷入局部最优，能更好地提高算法精度。但是 GA 算法具有交叉、变异操作，使得计算过程复杂度较高。

2.3 PSO-BP 神经网络

粒子群优化算法(PSO)是由 Kennedy 和 Eberhart[7] 发现的一种源于动物群体智能行为的非线性优化算法。通过个体极值(粒子的最优解)与全体极值(种群最优解)进行更新。

与其他优化算法相比，PSO 保留了全局搜索策略，使遗传操作更加简单，具有更快的计算速度和更好的全局搜索能力。和 GA 算法相比，PSO 算法比 GA 算法的规则更为简单，仅根据自己的速度来决定搜索，没有 GA 的明显的交叉和变异，仅通过当前得到的最优值来寻找全局最优，即操作简单。除此之外，GA 算法除了最后一代个体的信息外不能保存前期的迭代信息，而 PSO 可以给出多个迭代过程中的信息。

粒子群优化 BP 神经网络(PSO-BP)的步骤如下：

(1)初始化粒子群，包括：群体规模 N，每个粒子的位置 X 和速度 V。

(2)计算每个粒子的适应度值 Fit。

(3)比较每个粒子的 Fit、P_{best}，g_{best} 如果 Fit＞P_{best} 或 Fit＞g_{best}，则用 Fit 替换掉 P_{best} 或 g_{best}。

(4)按照公式更新粒子的位置和速度：

①$v[]=w*v[]+c1*\text{rand}()*([]-\text{present}[])+*\text{rand}()*([]-\text{present}[])$

②$\text{present}[]=\text{present}[]+v[]$

(6)当满足误差范围时退出优化，否则重新计算 Fit。

(其中 P_{best}：个体极值；g_{best} 全局极值)

3 三种回归预测实验

3.1 BP 神经网络

对于星系中的早型星系样本分为训练数据和测试数据，其中训练数据占样本总数 60%测试数据占样本总数的 40%。实验中通过建立三层的 BP 神经网络，输入层包括星系样本

u，g，r，i，z 五个波段的测光值以及 10 个颜色特征；隐含层神经元个数为 19；将估计值作为输出层。最后利用星系的光谱红移作为期望值，评判回归预测网络的准确性。BP 算法的实验结果如图 2 所示。

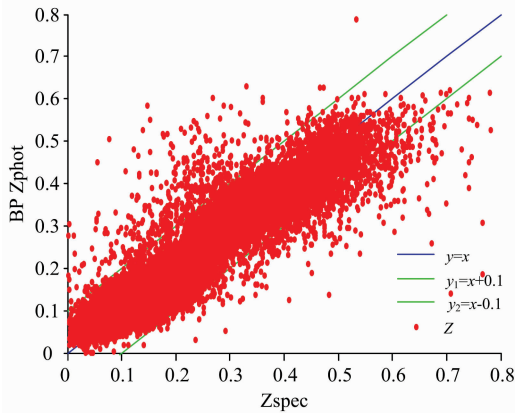

图 2　BP 算法的红移回归预测结果

Fig. 2　The redshift prediction results of the BP algorithm

由图 2 可知，BP 神经网络模型能够对早型星系的红移进行回归预测，模型在 $z<0.6$ 的低红移部分离散程度较小，能够达到很好的回归效果，$z>0.6$ 的高红移部分分散程度比较大，导致回归误差稍大，这也为后续的优化提供了思路。

3.2　GA- BP 神经网络

使用的数据源自 BP 神经网络回归预测实验。使用 GA 进一步优化 BP 算法，其中 GA 种群规模设为 190，迭代 100 次。通过遗传算法的选择、交叉、变异等一系列操作，不断更新权值和阈值，直至达到终止条件为止。GA-BP 算法的实验结果如图 3 所示。

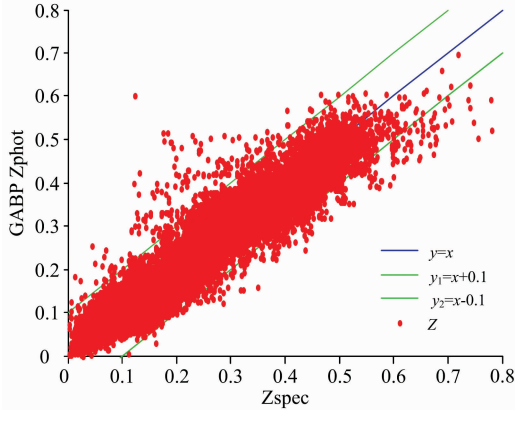

图 3　GA-BP 算法的红移回归预测结果

Fig. 3　The redshift prediction results of the GA-BP algorithm

由图 3 可知，GA-BP 网络对早型星系的预测结果具有改进作用，且早型星系在红移边界部分所出现的扁平化离散问题也有了很大的收敛效果。

3.3 PSO-BP 神经网络

PSO-BP 实验中使用的数据同 BP 神经网络回归预测实验中的数据。且 PSO-BP 实验中各参数设定：

①粒子群中的粒子数定为 100。

②学习因子 C1 和 C2 统一定为 2。

③粒子的宽度范围为 0.1。

④V_{max}：V_{max}过高，粒子可飞越最优解，V_{max}过低，粒子的局部最佳范围变小，将导致算法转化为部分最优值。故设定最 $V_{max}=10$。

⑤惯性权重：可以增强 PSO 的整体搜索能力，经过多次测试，本文中的惯性权重值为 0.729。

实验进一步使用粒子群算法对神经网络的权值和阈值进行全局寻优，将各层的连接权值编码成粒子，将输出均方误差作为系统的适应度值，在满足实验要求的迭代范围内，通过不断的迭代过程寻求网络的最优权值。PSO-BP 算法的回归预测结果如图 4 所示。

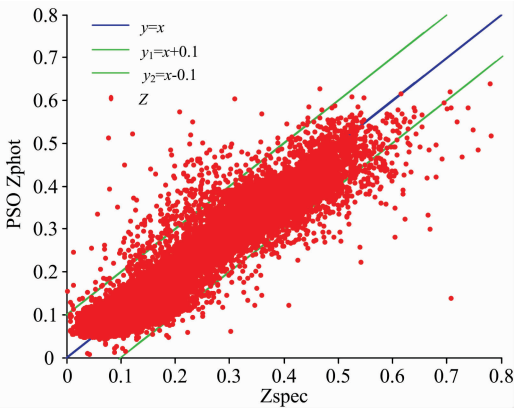

图 4　PSO-BP 算法的红移回归预测结果

Fig. 4　The redshift prediction results of the PSO-BP algorithm

由图 4 可知，PSO-BP 网络在早型星系的预测实验中，基本上能够很好地达到测光红移的预测效果。对早型星系的预测结果起到了改进作用，且相比较 BP 与 GA-BP 网络，PSO-BP 网络的收敛程度较优。

4　结果与讨论

为分析 PSO 对于 BP 神经网络的改进效果，将 PSO-BP 网络的回归预测结果分别与 BP 回归预测、GA-BP 回归预测的实验结果进行误差对比。采用均方误差(MSE)和均方根误差(RMSE)作为分析指标。比较结果见表 1。

(1)均方误差 MSE：

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(Z_{\mathrm{photo}}(i)-Z_{\mathrm{spec}}(i))^2$$

(2)均方根误差 RMSE

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z_{\mathrm{photo}}(i)-Z_{\mathrm{spec}}(i))^2}$$

从表 1 中可明显看出 PSO-BP 神经网络在早型星系的测光红移估计中，在精度上优于仅用 BP 神经网络的红移预测，并且也优于GA-BP神经网络的红移预测。在$Z<0.8$范围

表 1 BP 神经网络、GA-BP 神经网络、PSO-BP 神经网络测光红移误差比较

Table 1 Comparison of photometric redshift error estimation between BP neural network, GA-BP neural network, PSO-BP neural network

Error parameter	BP Early type Galaxy	GA-BP Early type Galaxy	PSO-BP Early type Galaxy
MSE	0.001 92	0.001 728	0.001 708
RMSE	0.043 38	0.041 750	0.041 330

内，PSO-BP 网络对早型星系红移预测的 MSE 值为 0.001 708，较 BP 网络提高了 11.1%，较 GA-BP 网络提高了 1.2%。PSO-BP 网络对早型星系红移预测的 RMSE 值为 0.04133，较 BP 网络提高了 5.8%，较 GA-BP 网络提高了 1.1%。将本实验结果与 Beck 文中的 K 近邻算法[9]预测星系测光红移的结果进行比较，在同样的样本集情况下，预测精度达到了相当的结果，并且 PSO-BP 的实现效率要高于 K 近邻回归的效率。经过 PSO 优化后的 BP 神经网络进一步减少了迭代次数，大大提高了算法的效率，不仅在效率上较其他算法有明显优势，而且在精度上也能保证很好的回归预测效果。

5 结 论

星系的测光红移数据量较大，对红移研究过程中使用的算法精度和效率有较高的要求。本研究主要对聚类后占比大的早型星系，采用 PSO-BP 算法对早型星系样本进行了测光红移回归预测，在 $z<0.6$ 的低红移部分能够很好地获得回归效果，并将结果与仅用 BP 算法和使用 GA-BP 算法的回归预测结果进行比较。经过对实验结果的分析，将 PSO 算法与 BP 算法相结合，充分利用 PSO 算法的优点，使得早型星系的回归预测收敛性更高，而且实验误差得到了改善，学习、泛化能力在一定程度上得到了提高。使得 PSO-BP 算法在回归预测精度上，不仅优于 BP 算法，而且也优于 GA-BP 算法。此外，PSO-BP 解决了 BP 神经网络易陷入局部最优的缺陷，也简化了 GA-BP 优化过程中的复杂操作，不仅在回归预测精度上有了提高，而且在算法效率上也得到了改善。

References

[1] Cohen S H, Kim H, Petty S M, et al. Pixel-by-Pixel SED Fitting of Intermediate Redshift Galaxies (C), 2015. 225.

[2] Kügler S D, Gianniotis N, Kai L P. A Spectral Model for Multimodal Redshift Estimation (C). Computational Intelligence, IEEE Xplore, 2016. 1.

[3] Zheng H, Zhang Y. Review of Techniques for Photometric Redshift Estimation. International Society for Optics and Photonics, Software and Cyberinfrastructure for Astronomy Ⅱ, 2012. 8451.

[4] Franco D Albareti, Carlos Allende Prieto, Andres Almeida, et al. The Astrophysical Journal Supplement Series, 2017, 233(2): 25.

[5] Robert Beck, Laszl Dobos, Tamas Budavari, et al. Monthly Notices of the Royal Astronomical Society, 2016, 460(2): 1371.

[6] Almosallam I A, Jarvis M J, Roberts S J. Monthly Notices of the Royal Astronomical Society, 2016, 462(1): 726.

[7] Cavuoti S, Amaro V, Brescia M, et al. Monthly Notices of the Royal Astronomical Society, 2017, 465(2): 1959.

[8] Shaun Cole, Steve Hatton, David H Weinberg, et al. Monthly Notices of the Royal Astronomical Society, 2013, 300(4): 945.

[9] Beck R, Dobos L, Budavári T, et al. Monthly Notices of the Royal Astronomical Society, 2016, 460(2): 1371.

[10] Kadim Taşdemir, Pavel Milenov, Brooke Tapsall. Computers & Electronics in Agriculture, 2012, 79(1): 92.

[11] Yao H Q, Jiang Y L. Applied Mechanics & Materials, 2014, 584-586: 1346.

[12] Kennedy J, Eberhart R. IEEE International Conference on (IEEE), 1995, 4: 1942.

Regression Prediction of Photometric Redshift Based on Particle Warm Optimization Neural Network Algorithm

MU Yong-huan[1], QIU Bo[1*], WEI Shi-ya[1], SONG Tao[1], ZHENG Zi-peng[1], GUO Ping[2*]

1. School of Electronics and Information Engineering, Hebei University of Technology, Tianjin 300400, China

2. School of Systems Science of Beijing Normal University, Beijing 100875, China

Abstract In addition to the spectral redshift of galaxies, the prediction of galaxies redshift has important research significance for studying the large-scale structure and evolution of the universe. In this paper, we use the metering and spectral data of 150 000 galaxies of SDSS DR13 released by the Sloan Sky Survey project to analyze the galaxies according to the color characteristics and clustering methods. The classification results show that the early galaxies account for a large proportion. In this paper, three different machine learning algorithms are compared to measure the redshift regression prediction of early galaxies and find the optimal method. In the experiment, the photometric values of the galaxy samples u, g, r, i, z and the 10 color features

obtained by the difference between the two bands are used as input data. First, the BP network is constructed, and the BP algorithm is used to measure the galaxies redshift. Then the Genetic Algorithm (GA) is used to optimize the parameters of the BP network, and the optimized GA-BP algorithm is applied to the regression prediction experiment of the early galaxies; considering the complex operation of the GA algorithm will affect the prediction efficiency. Moreover, the Particle Swarm Optimization algorithm not only has high stability and simple operation, so the Particle Swarm Optimization algorithm is used to optimize the BP network (PSO-BP) and Particle Swarm Optimization is used to optimize BP network (PSO-BP). By adjusting the weight method to improve the prediction efficiency and increase the stability, the particle swarm optimization algorithm is used to predict the redshift of the early galaxies in the galaxy samples. In the experiment, the spectral redshift is taken as the expected value, and the mean square error (MSE) is used as the error analysis index to judge the accuracy of the three algorithms. The PSO-BP regression prediction results are compared with the BP network model and the GA-BP network model. The experimental results show that the MSE value of the BP network is 0.001 92, the MSE value of the GA-BP network is 0.001 728, and the MSE value of the PSO-BP network is 0.001 708. The experimental results show that the PSO-BP optimization model used in this paper is superior to the BP neural network model and the GA-BP neural network model in terms of accuracy, which is respectively improved by 11.1% and 1.2%. It is superior to the traditional K-nearest neighbor test in efficiency, which overcomes the shortcomings of traversing all data samples in KNN algorithm and its generalization performance is better than that of other BP network optimization models.

Keywords Photometric redshift; Particle swarm optimization; PSO-BP optimization network; BP neural network; GA-BP neural network

(Received Jul. 19, 2018; accepted Nov. 25, 2018)

* Corresponding authors