计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):11−18 [doi: 10.15888/j.cnki.csa.007380]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

交叉验证的 BP 神经网络恒星光谱分类①

刘曼云[1], 赵正旭[2], 郭　阳[1], 赵士伟[1], 曹子腾[1]

[1](石家庄铁道大学 复杂网络与可视化研究所, 石家庄 050043)
[2](青岛理工大学 机械与汽车工程学院, 青岛 266520)
通讯作者: 刘曼云, E-mail: 1030462138@qq.com

摘　要: LAMOST 作为国家重大科学工程项目, 目前在世界上对光谱的观测、获取率最高, 为天文学的研究与发展提供大量的数据和信息资源. 根据 LAMOST 发布的恒星光谱数据文件, 从中提取出关于恒星光谱波长的数据信息, 对数据进行噪声剔除、数据降维、数据规范化、数据降维处理. 利用 BP 神经网络算法对数据进行分类处理, 根据分类结果正确率来判断 BP 神经网络模型的优劣. 但是 BP 神经网络对测试集数据的测试效果并不代表对其他数据具有同样的测试效果而且易产生过拟合, 所以采用交叉验证与 BP 神经网络相结合的方法, BP 神经网络算法可对多组不同的数据进行测试, 得到多组测试结果并求得平均值, 可得到 BP 神经网络模型相对稳定的测试结果并降低结果的随机性.

关键词: LAMOST; 光谱数据; 恒星光谱分类; 交叉验证; BP 神经网络

引用格式: 刘曼云,赵正旭,郭阳,赵士伟,曹子腾.交叉验证的 BP 神经网络恒星光谱分类.计算机系统应用,2020,29(5):11–18. http://www.c-s-a.org.cn/1003-3254/7380.html

Cross-Validation BP Neural Network Stellar Spectral Classification

LIU Man-Yun[1], ZHAO Zheng-Xu[2], GUO Yang[1], ZHAO Shi-Wei[1], CAO Zi-Teng[1]

[1](Institute of Complex Networks and Visualisations, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)
[2](School of Mechanical and Automotive Engineering, Qingdao University of Technology, Qingdao 266520, China)

Abstract: As a national major scientific engineering project, LAMOST currently has the highest observation and acquisition rate of the spectrum in the world, and provides a large amount of data and information resources for the research and development of astronomy. According to the stellar spectral data file released by LAMOST, the data about the wavelength of the stellar spectrum is extracted, and the data is subjected to noise culling, data dimensionality reduction, data normalization, and data dimensionality reduction processing. The BP neural network algorithm is used to classify the data, and the pros and cons of the BP neural network model are judged according to the correct rate of the classification results. However, the BP neural network test results of the test set data do not mean that it has the same test effect on other data and is easy to produce over-fitting, so the method of cross-validation combined with BP neural network is adopted. The BP neural network algorithm can test multiple sets of different data, obtain multiple sets of test results and obtain the average value, and obtain the relatively stable test results of the BP neural network model and reduce the randomness of the results.

Key words: LAMOST; spectral data; stellar spectrum classification; cross-validation; BP neural network

① 基金项目: 河北省自然科学基金 (F2018210058)
Foundation item: Natural Science Foundation of Hebei Province (F2018210058)
收稿时间: 2019-09-27; 修改时间: 2019-10-22, 2019-10-30; 采用时间: 2019-11-05; csa 在线出版时间: 2020-05-07

1 研究背景介绍

1.1 研究背景

宇宙之中充满了神秘与未知, 偌大的宇宙是如何形成以及演化的? 宇宙中无人知晓其数量的星系是如按照何种方式分布的? 宇宙中所有的星球、星体是如何在其轨道上有条不紊得运行的? 星系的光谱有哪些物理性质以及化学成分? 这些问题都激发着我们对天文学的探求欲. 天体光谱是对天文学的探索与认知的主要的资源信息, 天体光谱的形状与星体的物理性质、化学成分、运动轨迹、运动状态有着密切的关系, 光谱数据对发现恒星的特征信息提供有利的帮助. 星体会抛射出星周气体, 星周气体具有不同的状态、性质, 其不同的特征会造成恒星光谱发射线在轮廓、宽度上的不同, 很少的星体会存在发射线, 发射线表示这些星体经历或正正经历着一些不稳定的变化, 这类星体的发现可以用于研究星体的演化过程.

美国数字巡天计划 (SDSS)、澳大利亚英澳天文台[1,2]等对于天体中数量庞大的恒星、星系、类星体的光谱信息进行的获取, 为天文学的发展提供了巨大的信息资源. 虽然光谱样本的数据量很大, 但仍不能满足科学界对光谱信息的需求, 大天区面积多目标光纤光谱天文望远镜, 简称为 LAMOST (Large Sky Area Multi-Object Fibre Spectroscopy Telescope), 是我国自主设计并研制的反射施密特望远镜, 就是为了满足获取海量光谱的要求而建造, 它位于国家天文台的兴隆观测站.

1.2 LAMOST 简介

LAMOST 的大口径为 4 米, 大视场为 5 度, 光谱观测系统包括 4000 根光纤[3], 连接着 16 台光谱仪, 其设计上的创新为我国的天文学带来飞跃式的发展. LAMOST 与国际上同类型的望远镜在口径大小、观测范围等方面相比都更胜一筹, 是美国数字巡天计划 (SDSS)、澳大利亚英澳天文台巡天的观测效率的 10~15 倍[4]. LAMSOT 的建造的第一个科学目标是根据 LAMOST 所观测到的星体信息, 可对宇宙和星系进行研究, 通过所获取的星系光谱得到星系的红移数据, 根据红移可得到其距离, 有了距离便有了其三维分布, 从而知晓整个宇宙空间的结构. 第二个科学目标是通过对更暗的恒星的观测, 掌握银河系中更远处的恒星的分布以及运行的情况, 从而对银河系的结构进行了解. 第三个科学目标是通过对大量天体的光谱观测, 根据光谱中包含的不同射线来进行各类天体多波段交叉证认. 有关光谱的理论很多, 说法不一, 光谱信息还能对大量的光谱理论的正确性进行验证[5].

1.3 研究意义

LAMOST 所观测到的海量光谱信息构成了一个庞大的天文数据库, 是天文界重要的信息资源, 在如今的大数据时代和经济建设时代, 信息资源对国家的发展具有战略性意义[6]. LAMOST 所发布的信息资源为 FITS 格式文件, FITS 文件中含有光谱的诸多属性信息, 可以从 FITS 文件中提取出与研究相关的属性信息, 通过数据挖掘等技术可对光谱数据之间的联系进行分析、研究, 发掘出恒星光谱分类更准确的依据, 还有助于发现未知天体以及脉冲星, 脉冲星的信号可为星际导航提供依据.

面对 LAMOST 所发布的海量的恒星光谱数据, 传统的人工处理恒星光谱的方法已经不再适用, 需要找到更高效的、可利用计算机来处理的恒星光谱分类算法, 而数据挖掘中的分类算法就能很好的应用于此. 根据适合的光谱自动分类算法处理大量的恒星光谱数据, 提高恒星光谱分类速度、效率, 高效且便捷地应用于大量数据的分析、处理.

2 数据预处理

2.1 数据噪声剔除

LAMOST 发布的光谱信息包括恒星、星系、类星体、未知天体的光谱信息, 将下载的恒星光谱数据进行读取, 并整合到一个 csv 文件中, 其中可能会掺杂着其他星体的光谱数据, 要将其他星体的光谱数据进行删除, 具体方法为将文件中的数据根据 class 的属性值进行排序, 然后将 class 的类别不为 star 的数据选中并删除; 恒星光谱中又包含着诸多子类, 最常见的恒星光谱分类系统是哈佛系统[7], 此方法为美国一天文学家的创新, 将恒星按照温度的高低进行排序, 恒星光谱按照温度的降序排列为: O, B, A, F, G, K, M, R, S, N 几个大的类别, 每个大类别下又分为诸多小类别, 用十进制数值表示, 例如光谱 A 类型下又分为 A0, A1, A2, ⋯, A9 等 10 个子类型, 这为恒星光谱的分类方面做出了创新性的贡献, FITS 文件中有一个关于光谱的属性名称为 subclass, 含义为恒星光谱的子分类, 需要把子分

类的值不是正确范围内以及类别不明确的光谱数据根据同样的方法进行删除.

2.2 数据降维

在收集、处理数据的过程中, 经常会遇到特征维度很多的数据, 高维度的数据会使我们在后续对数据的处理上非常繁琐、困难, 而且有些数据含有冗余的信息、噪声信息等, 这些数据会对结果的正确性造成影响. 所以, 需要通过剔除数据中的冗余信息、降低数据维度来达到降低数据量的目的, 但又对数据中的重要信息有所保留. 降维是数据预处理中非常重要的一个环节, 如果不对数据进行预处理, 在很多数据处理的算法中便很难得到理想的结果.

在对数据进行降维所选择的方法为 PCA 降维. PCA 是降维方法中使用最频繁的降维方法, PCA 通过线性投影将高维数据在某方向上进行投影, 投影到低维的空间中, 此方向可使在投影的维度上, 数据之间具有最大的方差, 这样就能保证得到的数据丢失原始数据最少, 保留了原始数据中更多的数据特点. PCA 的降维过程, 更具体的来说, 有一个具有 n 维属性特征的数据集, 数据集中有 m 个样本点, PCA 需要做的就是将具有 n 维属性特征的数据集降维到 $n'(n'<n)$ 维, 降维后的样本点之间具有最大的方差, 所以降维后得到的 n'维数据集保留了原始数据集中最大的差异性, 将数据损失性降到最低. 本实验使用 Python 来对数据进行降维, skleran 中已经有实现 PCA 成熟的包, 通过 from sklearn.decomposition import PCA 来导入 PCA 包, 所以在数据降维的过程中直接调用即可.

核心代码如算法 1.

算法 1. PCA 降维算法

```
pca=PCA(n_components=22) #n_components 为最终得到的维度
newDataMat = pca.fit_transform(DataMat) #DataMat 为原始数据,
newDataMat 为降维#后的数据
print(newDataMat) # 输出降维后的数据
```

2.3 数据规范化

由于天体距离地球的远近、天体的亮度不尽相同, 所以 LAMOST 观测到的光谱数据的数量级可能会存在着很大的差异[4], 对数据进行归一化处理可使数据处于同一数量级, 从而也消除了奇异样本数据对数据处理结果导致的不良影响. 数据归一化过程是将数据进行同比例缩放, 对每个数据进行同样的处理, 使每个数据在被处理后落入一个特定的范围, 如范围可为[–1, 1]或者[0, 1]等. 在对数据进行分类时, 如果分类算法涉及到了神经网络、最近邻分类和聚类等, 在对数据进行预处理的过程中实现对数据的归一化操作, 会对后续的分类挖掘提供很大的帮助.

对数据进行规范化采用的方法是最小-最大规范化, 此方法的思想是对原始数据进行线性变换, 将结果变换到[0, 1]范围内, 并且不对原始数据之间的关系进行改变. 对原始数据中的每个数值进行如下变换:

$$x' = \frac{(x-min)}{max-min}(new\_max - new\_min) + new\_min \quad (1)$$

其中, max、min 分别为原始样本数据中的最大值、最小值, new_max、new_min 分别为原始数据经过规范化后的数值中的最大值、最小值, 比如, new_max、new_min 经常取到的值为 1.0, 0.0, 此方法可以根据数据特点来指定数据进行归一化后的取值范围.

2.4 数据平衡

对数据进行分类, 根据其类别保证数据量的同时, 也要保证不同类别的数据量的均衡, 如果不同类别的数据量相差较大, 在对数据进行分类算法处理的过程中, 分类算法会对数量较大的类别投入更多的关注, 而忽略数量较小的类别, 导致训练出的分类器的分类性能很低.

重采样是通过对数量较少的样本进行增加, 减少数量较多的样本来使不平衡的数据集变得更加平衡, 从而改善分类算法的结果. 欠采样也称为下采样, 是处理数量大的类别所用到的方法, 可通过丢弃数量大的类别中的部分数据、去除冗余样本数据、删除边界样本数据等方法, 其目的是减少多数类别的数据量来保证不同类别的数据集的平衡. 欠采样在对数量大的类别样本进行去除的过程中, 会去除掉具有重要信息的样本数据, 而若只去除具有错误数据的少数样本, 又不能达到平衡数据集的目的. 过采样也称为上采样, 过采样是通过对数量少的类别的数据进行复制, 来提高少数类别的数据量, 但如果某类别的数据量非常小, 而进行过采样后所要求的数据量又较多, 在对此类别数据进行过采样后, 很容易导致过拟合, 对数据量小的类别的识别率会没有意义. 为防止过多相同数据的生成, 可采用 SMOTE 算法, 此算法同样可对数据量小的数据

进行过采样, 首先为每个少数类样本选出几个近邻的样本, 将每个样本与其近邻的几个样本相连接, 然后在连接线上随机取点作为新生成的样本, 这些新样本与原始样本并不相同, 从而扩大数据量少的类别, 使得不同类别的恒星光谱具有大致相同的数据量. SMOTE 算法也是可以直接调用 SMOTE 库来实现数据的平衡.

核心代码如算法 2.

算法 2. 过采样 SMOTE 算法

```
from imblearn.over_sampling import SMOTE #导入 SMOTE 库
print(sorted(Counter(y_train).items()) #查看#原始数据中不同类别恒星光谱的数据量
X_train,Y_train=SMOTE().fit_sample(x_train,y_train)
#对原始数据进行过采样处理
print(sorted(Counter(y_train).items()) #得#到不同类别的恒星光谱具有相同的数据量
```

对数据的处理完成后, 就需要采用数据挖掘中的分类算法对数据进行分类. 之所以采用 BP 神经网络算法, 是因为其具有其他算法所不具有的优点. BP 神经网络实现了输入数据经过网络的传输, 最终输出的过程, 在这个过程中, BP 神经网络能够以任意精度逼近非线性连续函数, 具有很强的非线性映射能力; BP 神经网络在训练数据的过程中, 能够提取出输入数据与输出数据之间的“规则”, 并自适应的将“规则”记忆于网络的权值之中, 所以 BP 神经网络具有自适应能力; BP 神经网络具有着较为复杂的网络结构, 当其中的部分神经元受损后, 不会对全局的 BP 神经网络结构造成太大的影响, 甚至仍然可以正常工作, 具有一定的容错能力. BP 神经网络算法适合处理内部机制复杂的问题, 而且具有较强的自适应能力与容错能力, 而恒星光谱数据量大, 数据关系复杂, 所以采用 BP 神经网络算法对恒星光谱数据进行分类是很合适的选择.

3 BP 神经网络算法研究

3.1 BP 神经网络原理

神经网络, 顾名思义, 模仿人脑中的传输过程, 首先是输入信号进行输入, 经过神经元的层层传输以及处理, 最终输出反馈, 即得到输出结果. BP 神经网络是一种包括 3 层或 3 层以上的阶层型神经网络, 标准的 BP 神经网络模型有 3 层, 包括输入层、隐含层、输出层, 如图 1 所示. BP 神经网络具有一层输入层, 一层输出层, 可包含多层隐含层, 具有很强的记忆与泛化能力[8], 通过对网络中的权值和阈值不断地进行调整, 来降低预测误差平方和.

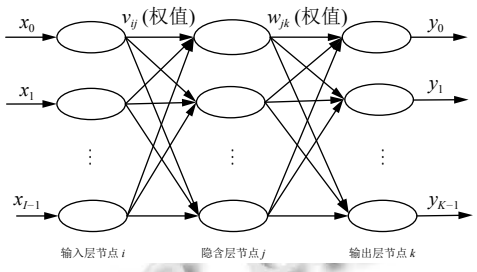

图 1 3 层 BP 神经网络结构示意图

利用 BP 神经网络系统实现对恒星的分类, 其过程类似于上述中信号在神经网络中的传输过程. 可以把目标源的轮廓信息、颜色信息、光谱能量分布信息作为输入数据, 输入到神经网络系统的输入层, 神经网络系统按照一定的规则、特定函数来对这些信息进行判断、处理、分析, 输出层就会给出目标源信息的特征量, 然后分析、总结输出层输出的数据信息, 判断、总结目标源的特点, 从而可以实现对目标源的分类. BP 神经网络包括两个阶段, 第一个阶段称为正向传播过程, 输入数据从输入层输入神经网络, 经过层层网络的计算、处理, 得到输出层各个单元的输出值. 记 BP 神经网络的输入层节点数为 I, 输入向量记为 $X_p=(x_0, x_1, x_2, \cdots, x_{I-1})$. 输出层的节点数为 K, 隐含层节点数为 J, 输入层节点数一般表示数据集中对于数据的属性描述有 I 个, 将属性描述对应的属性值作为 BP 神经网络的输入值, 而数据集中每个样本的分类类别作为 BP 神经网络的期望输出值, 输出值为 K 个, 隐含层第 j 个神经元的阈值 α_j 来表示, 输出层第 k 个神经元的阈值用 β_k 来表示, 输入层第 i 个神经元与隐含层第 j 个神经元间的权值用 v_{ij} 来表示, 隐含层第 j 个神经元与输出层第 k 个神经元间的权值用 w_{jk} 来表示. 对于 BP 神经网络, 总会有一个理想化的输出向量, 称为期望输出向量, 表示为 $D_p=(d_1, d_2, d_3, \cdots, d_{K-1})$

输入层数值表示:

$$O_i = x_i,\ \ i = 0,1,2,\cdots,I-1 \tag{2}$$

输入层的每个神经元的输入值分别与其连接的隐含层的第 j 个神经元对应的权值相乘, 然后求和, 得到隐含层中第 j 个神经元的输入值, 表示为:

$$net_j = \sum_{i=0}^{I} v_{ij} O_i,\ \ i = 0,1,2,\cdots,I-1 \tag{3}$$

隐含层的输出: 隐含层的第 j 个神经元输入值 net_j 经过激活函数的处理得到其输出值:

$$O_j = f\left(net_j\right),\ \ j = 0,1,2,\cdots,J-1 \tag{4}$$

其中, 激活函数一般选取 Sigmoid 函数[9]:

$$f(net) = \frac{1}{1+e^{-net}} \tag{5}$$

隐含层的每个神经元的输出值分别与其连接的输出层的第 k 个神经元对应的权值相乘, 然后求和, 作为输出层中第 k 个神经元的输入值, 表示为:

$$net_k = \sum_{j=0}^{J} w_{jk} O_j,\ \ i,j = 0,1,2,\cdots,J-1 \tag{6}$$

输出层的输出: 输出层的第 k 个神经元输入值 net_k 经过激活函数的处理得到其输出值:

$$O_k = f(net_k),\ \ k = 0,1,2,\cdots,K-1 \tag{7}$$

BP 神经网络的实际输出与期望输出之间总会有一定的差别, 把差别称为误差, 对于每个样本的误差函数的计算公式为:

$$E_p = 1/2 \sum_{k=0}^{K-1} \left(d_k^p - O_k^p\right)^2 \tag{8}$$

如果数据集中有 N 个样本, 则 N 个样本的总误差计算公式为:

$$E = \frac{1}{2N} \sum_{p=0}^{N-1} \sum_{k=0}^{K-1} \left(d_k^p - O_k^p\right)^2 \tag{9}$$

其中, E_p 表示 p 的输出误差, d_k^p 表示样本 p 的期望输出, O_k^p 表示 BP 神经网络的实际输出.

第二个阶段为反向传播过程, 经输出层输出的数据得到输出误差, 输出误差反向向前传输经过各隐含层, 计算隐含层各单元的误差, 再根据此误差来对前层的权值进行修正. 具体过程为:

首先对 BP 神经网络中的输出层与隐含层之间的连接权值进行调整, 输出层的误差表示为[10]:

$$\delta_k = (d_k - O_k) f'(net_k) \tag{10}$$

对于输出层与隐含层之间的权值调整为[10]:

$$\Delta w_{jk} = -\eta O_j (d_k - O_k) O_k (1 - O_k) \tag{11}$$

其中, η 为学习步长, 取值区间为 (0,1).

然后对 BP 神经网络中的隐含层与输入层之间的连接权值进行调整, 隐含层的误差表示为[10]:

$$\delta_j = f'\left(net_j\right) \sum_{k=0}^{K-1} \delta_k w_{jk} \tag{12}$$

对于隐含层与输入层之间的权值调整为[10]:

$$\Delta v_{ij} = \eta O_j \left(1 - O_j\right) \sum_{k=0}^{K-1} \delta_k w_{jk} O_i \tag{13}$$

BP 神经网络算法的流程图如图 2 所示.

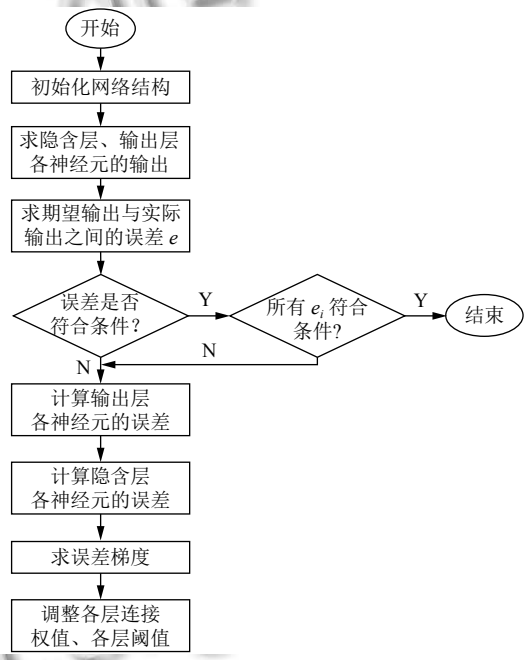

图 2 BP 神经网络算法流程图

3.2 BP 神经网络实现

读取存储恒星光谱的光谱文件, 提取出在不同波长处的光谱波长信息以及光谱类别信息, 每个文件中共有 3909 个波长数据, 提取出光谱信息后, 将其中不是恒星以及恒星子类别不明确的光谱数据进行剔除, 将光谱波长数据中的冗余数据错误数据进行删除, 然后对数据进行降维操作, 选择将光谱波长数据降维成 22 维数据, 既提取出了其中的重要数据信息, 又减少了数据量, 方便对数据的操作处理. 然后对数据进行归一化处理, 消除数据之间的数量级差距, 同时也减小数据, 降低对数据的计算量. 还要对数据进行平衡处理, 通过采样操作使不同类别的恒星光谱数据具有相似的数

据量.

将恒星光谱的波长信息作为 BP 神经网络的输入, 类别信息作为 BP 神经网络的期望输出, 输入数据经过 BP 神经网络层层传递以及反馈, 经过 BP 神经网络权值、阈值的不断修正, 使输入数据经过 BP 神经网络的传输后, 实际的输出结果能最大限度的接近期望输出, 从而使误差达到最下, 即较高的分类正确率.

BP 神经网络隐含层的确定可根据输入层与输出层节点数目来确定, 假设 BP 神经网络的输入层节点为 m 个, 输出层节点为 n 个, 则隐含层节点的选取范围为$(m+n)^{\frac{1}{2}}+a$, $a\in[1,10)$, 当设置 BP 升级网络的输入层节点为 22, 隐含层节点为 13, 输出层为 5, 基于交叉验证的 BP 神经网络得到的结果有时为 70.34%, 而有时又可达 76.66%, 其中的正确率还是有一定差距的, 因为每次选取的训练集、测试集是不同的, 所以导致最终的预测结果也是不同的.

4 基于交叉验证的 BP 神经网络

4.1 交叉验证

交叉验证是一种估计泛化误差的模型选择方法, 在进行误差估计之前不需要任何假设条件, 操作简单、方便, 交叉验证在各种类型的模型选择中都可适用, 所以其应用非常广泛[11]. 当分类算法在对数据中的训练集进行训练后得到训练模型, 如果用此训练模型再去对与训练集有交集数据的测试集进行误差估计, 会导致预测误差非常低, 得到错误的预测结果.

经常用的交叉验证方法为 K-折交叉验证方法 (K-folder Cross Validation, K-CV). K-折交叉验证是将数据集随机分成 k 份, 其中 k–1 份作为训练集, 1 份作为测试集, 在训练的过程中, 依次从 k 份中选择一份作为测试集, 剩余的 k–1 份为训练集, 每次用训练集进行模型训练, 用训练出的模型对测试集进行测试, 进行结果预测, 得到预测的正确率与误差[12]. K 份数据就需要进行 k 次训练与测试, 最后将得到的 k 个结果求平均值作为最终预测结果. K-折交叉验证的优点是数据集中的所有数据都作为了训练集和测试集, 每个数据样本都被验证过, 这样可使预测结果更具有客观性, 对预测结果求平均值也降低了偶然性误差.

在数据需要进行过采样时, 需要注意与交叉验证过程的顺序问题. 如果在交叉验证之前进行过采样同样会导致过拟合的问题, 因为在对数量小的类别进行过采样后, 对数据集进行随机划分, 不能保证测试集与训练集中的数据没有交集. 用一个例子来说明, 如图 3, 最左边一列为原始数据, 假设包含 3 个类别, 其中两个类别的数据量较小; 接下来如果首先对少数类别的数据进行过采样操作, 得到图中第 2 列数据集; 然后进行交叉验证中的数据划分过程, 会发现训练集与测试集中包含着同样的数据, 会导致最终结果产生过拟合的问题.

所以, 需要在交叉验证后对数据进行过采样操作, 如图 4 所示, 首先将验证样本从数据集中挑选出来, 剩余数据作为训练集; 然后对训练集中数量小的类别进行过采样, 得到如图中第 3 列所示, 此时训练集与测试集中的数据是没有重复的, 会防止预测结果产生过拟合. 在进行 10 折交叉验证的过程中, 同样是在交叉验证的每次循环中做过采样, 即每次在训练集中插入与测试集无交集的数据来保证数据平衡.

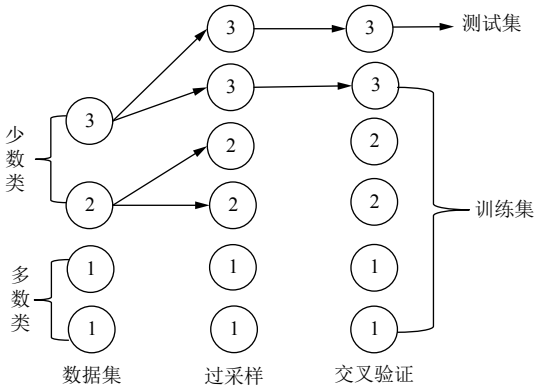

图 3 先过采样再交叉验证

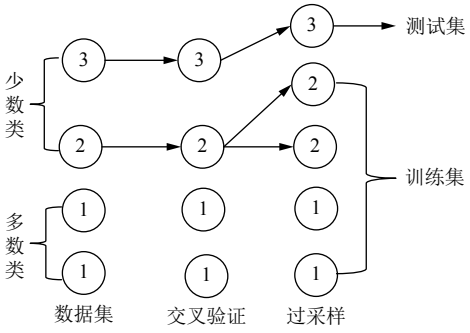

图 4 先过采样再交叉验证

4.2 基于交叉验证的 BP 神经网络

BP 神经网络中存储着诸多信息, 主要有两方面的信息, 一方面是网络的结构, 包括网络的输入层、隐含层以及输出层的节点个数以及隐含层的层数; 另一方面网络中连接每一层之间的权值. 网络中的输入层、输出层节点个数一般是可以确定的, 而隐含层层数与节点个数则由用户凭经验而设定, 设定的过小或过大都会造成不好的影响; 网络中的权值开始是在一个范围内进行初始化的, 在反向传播过程中, 会根据误差来进行调整. 用 BP 神经网络算法对数据进行训练、测试时, 测试的只是其中的一部分数据, 对这部分数据的测试结果并不代表对其他数据具有相似的测试结果, 而采用交叉验证的 BP 神经网络, 可使所有的数据都用于训练模型, 最终得到的结果会稳定、可靠; 实验过程中消除了数据分配的随机性而带来的误差影响, 确保实验的可重复性.

利用 10 折交叉验证进行 BP 神经网络, 将数据集平均分成 10 份, 每次从中选出 1 份作为测试集, 剩余 9 份作为训练集, BP 神经网络用训练集来训练出网络模型, 得到网络中的权值、阈值等, 确定了 BP 神经网络模型, 再用训练好的模型对测试集进行结果预测. 具体思想过程如图 5 所示, 将数据集平均分成 10 份, 分别用编号 1, 2, 3, …, 10 来表示, 第 1 次运行过程是将编号为 1 的那部分数据作为测试集, 剩余编号为 2~10 的 9 份数据作为训练集, 得到第 1 个测试结果 1, 第 2 次运行过程是将编号为 2 的那部分数据作为测试集, 剩余的编号为 1、3~10 的 9 份数据作为训练集, 得到测试结果 3, 以此类推, 最后一次训练过程是将编号为 10 的那部分数据作为测试集, 剩余的编号为 1~9 的 9 份数据作为训练集, 得到测试结果 10. 然后对 10 次的测试结果求平均值, 得多最终的测试结果.

4.3 实验结果

利用同样的数据作为实验的输入数据, 类别信息作为期望输出数据, 输入层节点为 22 个, 输出层节点为 5, 设置隐含层节点为 12, 基于交叉验证的 BP 神经网络得到的结果为: [72.92%, 79.12%, 79.17%, 67.7%, 65.6%, 78.1%, 80.2%, 69.8%, 74%, 76.7%]. 从 10 次预测的 10 个结果中可以发现, 最小的预测正确率为 65.6%, 而最大的预测正确率为 79.17%, 最大值与最小值之间还是有一定的差距的, 所以利用交叉验证, 进行 10 次训练与预测, 再求得 10 次预测结果的平均值, 会得到一个较为平稳、准确的数据结果为 74.331%. 对预测结果的影响因素有输入数据的质量、隐含层节点数的定义等, 所以在确定 BP 神经网络输入层节点、输出层节点后, 还需对隐含层节点的数量进行不同的设定来寻找较高的预测结果[13]. 基于大量实验, 发现当设定隐含层节点数为 14 时, 得到的较高的预测结果为: [82.24%, 78.5%, 80.37%, 83.18,%, 79%, 84.11%, 81.3%, 83.18%, 74.3%, 78.28%]. 从数据中同样可以发现, 最小值为 74.3%, 最大值为 84.11%, 最大值与最小值差距较大, 所以对 10 次预测结果求得均值为 80.446%.

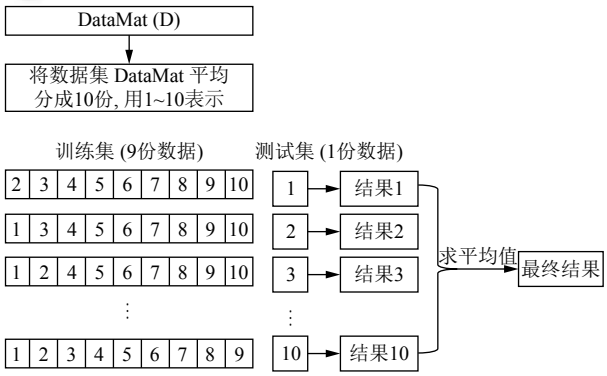

图 5 10 折交叉验证过程示意图

5 结语

在对数据进行分类前, 需要对数据进行噪声剔除、数据降维、数据规范化、数据平衡预处理, 可保证数据在训练过程中, 减少训练时间, 并且可使训练得到的 BP 神经网络更为准确, 从而可以得到较高的预测结果. BP 神经网络结构中的各层节点都是由数据来控制的, 但隐含层的节点的设定在一定范围内是随机的, 不同的隐含层节点的设定 BP 神经网络的预测结果是有一定的影响的, 所以需要基于大量的实验来确定使 BP 什么网络具有最高预测性能的隐含层节点数目. 基于交叉验证的 BP 神经网络, 几乎所有的数据都经过了 BP 神经网络算法的训练, 既可以在对数据进行过采样的过程中防止产生过拟合, 训练出的 BP 神经网络更为稳定, 降低了随机性, 而且多次训练得到的预测结果再求均值, 可以避免偶然的误差对结果造成的影响, 使测试结果更接近正确的、真实的结果.

参考文献

1 毕立鹏. LAMOST 低质量光谱交互式分析平台的设计与实现[硕士学位论文]. 济南: 山东大学, 2016.

2 赵永恒. 大规模天文光谱巡天. 中国科学: 物理学力学 天文学, 2014, 44(10): 1041–1048.

3 林雪梅. ANN 在天体光谱分类及恒星大气参数测量中的应用[硕士学位论文]. 济南: 山东大学, 2012.

4 艾丽雅. 天体光谱的分类算法研究[硕士学位论文]. 鞍山: 辽宁科技大学, 2016.

5 韦鹏. LAMOST 一维光谱自动处理[硕士学位论文]. 济南: 山东大学, 2011.

6 任利敬, 赵正旭, 陶智. 信息传承与长期保存技术策略研究. 兰台世界, 2016, (13): 25–27.

7 陈淑鑫, 孙伟民, 孔啸. LAMOST 恒星分类模板间相似性度量分析. 光谱学与光谱分析, 2018, 38(6): 1922–1925.

8 李老三, 辛军饷. 基于 BP 神经网络富水岩层围岩变形量预测. 重庆建筑, 2019, 18(4): 49–53. [doi: 10.3969/j.issn.1671-9107.2019.04.49]

9 贾伟, 赵雪芬. 改进量子粒子群 BP 神经网络参数优化及应用. 软件导刊, 2019, 18(10): 30–35.

10 王振武, 徐慧. 数据挖掘算法原理与实现. 北京: 清华大学出版社, 2015. 127–129.

11 阎少宏, 吴宇航. 基于交叉验证的级联 BP 神经网络的焦炭质量预测模型. 信息记录材料, 2018, 19(10): 223–224.

12 林悦, 夏厚培. 交叉验证的 GRNN 神经网络雷达目标识别方法研究. 现代防御技术, 2018, 46(4): 113–119. [doi: 10.3969/j.issn.1009-086x.2018.04.018]

13 丁常富, 王亮. 基于交叉验证法的 BP 神经网络在汽轮机故障诊断中的应用. 电力科学与工程, 2008, 24(3): 31–34. [doi: 10.3969/j.issn.1672-0792.2008.03.009]