

The Use of Support Vector Machines and Artificial Neural Networks to Fill Missing Data

Nan Zhang, Li Cheng, Peng Wang*

College of Sciences, Beijing Forestry University, Beijing

Email: 771568416@qq.com, 1790709839@qq.com, *wangpeng@bjfu.edu.cn

Received: Jul. 21st, 2017; accepted: Aug. 8th, 2017; published: Aug. 14th, 2017

Abstract

In this paper, we use the built-in data set iris to select the sample data according to the need to establish the learning samples, simulate the missing values of the biological sample attributes and the missing two types of missing data. Using MATLAB as the tool, we use the support vector machine and the artificial neural network to carry out the missing values to fill. For the case of missing attribute values in the biological sample data, the support vector machine and the artificial neural network can be used for the regression filling, and the applicability of the BP neural network and the RBF neural network can be compared. For the sample of the biological sample, missing situation, using support vector machine classification to fill. The results show that the adaptability of the number of hidden neurons in the RBF network is more stable than that of the BP network when the neural network is used to predict the missing attribute values. Compared with the artificial neural network, the support vector machine for the application, do not rely on the designer experience and has generalization ability.

Keywords

Biological Sample, Filling Missing Data, Support Vector Machine, Artificial Neural Network

利用支持向量机和人工神经网络填补缺失数据

张楠, 程理, 王鹏*

北京林业大学理学院, 北京

Email: 771568416@qq.com, 1790709839@qq.com, *wangpeng@bjfu.edu.cn

收稿日期: 2017年7月21日; 录用日期: 2017年8月8日; 发布日期: 2017年8月14日

*通讯作者。

摘要

本文从R内置数据集iris中按需要选取样本数据建立学习样本,模拟生物样本属性值缺失和类别缺失两种缺失数据的情况,以MATLAB为工具,利用支持向量机和人工神经网络对缺失值进行填补。对于生物样本数据中存在属性值缺失的情况,可以分别采用支持向量机和人工神经网络进行回归填补,并对BP神经网络和RBF神经网络的适用性进行了对比;对于生物样本数据中存在样本类别缺失的情况,采用支持向量机分类填补。结果显示,用神经网络预测填补缺失的属性值时,RBF网络对隐层神经元数目选取的自适应性使之比BP网络更为稳定;相比人工神经网络,支持向量机对有限的样本更为适用,并且不依赖设计者经验,泛化能力强。

关键词

生物样本, 填补缺失数据, 支持向量机, 人工神经网络

Copyright © 2017 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

一个简单的生命,也包含庞大的数据信息,数据构成十分复杂。随着生物科技的蓬勃发展,与生物有关的统计研究逐步展开,生物学数据的处理与分析成为一个重要的研究领域。有时我们获得的生物学样本数据不完整,导致研究工作困难重重。如何填补缺失的数据,也就成为了一个亟待解决的问题。

在生物样本数据中,通常情况下缺失的数据无非两种:属性值和分类值。如果缺失了属性值,可以进行回归填补;如果缺失了类别值,可以进行分类填补。这些都可以用支持向量机实现。就支持向量机而言,分类问题和回归问题都是根据训练样本找到一个实值函数 $g(x)$: 回归问题就是给定一个新的模式,根据训练样本确定一个实值函数 $g(x)$, 使用 $y = g(x)$ 推断任一输入 x 所对应的输出 y (实数); 分类问题就是给定一个新的模式,根据训练样本找到一个实值函数 $g(x)$, 使用 $y = \text{sign}(g(x))$ 推断任一输入 x 所对应的类别(如: +1, -1) [1]。文献[2]使用 RBF 神经网络对上证指数进行预测,类似地,填补缺失的属性值也可以用人工神经网络回归预测进行处理。本文从 R 内置数据集 iris 中按需要选取样本数据建立学习样本,模拟生物样本属性值缺失和类别缺失两种缺失数据的情况,以 MATLAB 为工具,采取支持向量机和人工神经网络的模型,以无缺失值的样本预测有缺失值的样本中的缺失值,从而进行填补。属性值缺失值用支持向量机或人工神经网络进行回归填补,样本类别缺失值用支持向量机分类填补。神经网络推导出的各种算法很难在样本数据有限时取得理想的应用效果,需要设计者有效利用自己的经验。与神经网络相比,支持向量机能够基于有限的样本信息求解,同时避免了神经网络实现中的经验成分。

2. 填补缺失的样本属性值

2.1. 支持向量机回归填补

支持向量机 SVM (Support Vector Machines)是由 Vanpik 领导的 AT&TBell 实验室研究小组在 1963 年提出的一种新的分类技术,在解决小样本、非线性、高维模式识别问题中表现出许多特有的优势,并且可以推广应用到函数拟合等问题中去[3]。

我们从 R 语言内置的 iris 数据集集中的 setosa 类别选取 20 个样本组成一个新的数据集, 作为网络学习样本, 并删去第二十个样本的最后一个属性值(Petal.Width), 模拟数据缺失的情况, 如表 1。其实, 无论缺失的是什么位置的数据, 只要将没有缺失数据的行列集合到一起, 作为训练集, 用来预测数据不完全样本的缺失属性的值。选取 1 到 19 个样本的 Sepal.Length、Sepal.Width、Petal.Length、Petal.Width 为自变量, 2 到 20 个样本的 Petal.Width 为因变量。

SVM 的实现使用 MATLAB 的 libsvm 工具箱, 实现数据归一化预处理, 寻找回归的最佳参数, 参数选择结果图(等高线图)如图 1。用找到的最佳参数对 SVM 进行训练, 再对原始数据进行回归预测, 得回归预测数据与原始数据对比图, 如图 2。SVM 回归预测的均方误差 $MSE = 0.0312355$, 相关系数 $R = 67.5382\%$, 缺失值填补为 0.2969。

2.2. 人工神经网络回归填补

人工神经网络(Artificial Neural Network, ANN)是由神经元相互连接, 通过模拟人脑神经处理信息的方式, 进行信息并行处理和非线性转换的复杂网络系统, 在控制与优化、预测与管理、模式识别与图像处理、通信等方面得到了十分广泛的应用[4]。前向反馈(Back Programming, BP)网络和径向基(Radical Basis Function, RBF)网络是目前应用最广泛的两种网络。

Table 1. A set of biological samples with missing attribute values

表 1. 有缺失属性值的生物样本集

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3	1.4	0.1
14	4.3	3	1.1	0.1
15	5.8	4	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	

注: 20 号样本缺失的 Petal.Width 真实值为 0.3。

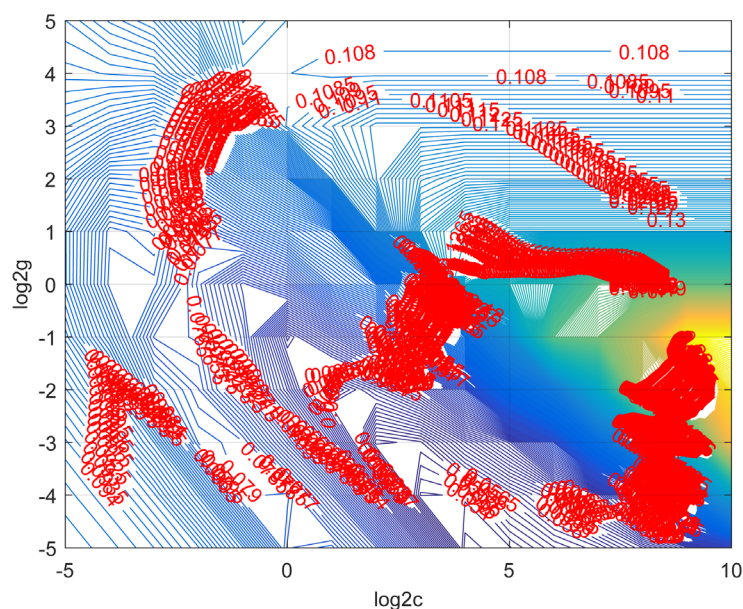


Figure 1. Parameter selection results (contour map)

图 1. 参数选择结果图(等高线图)

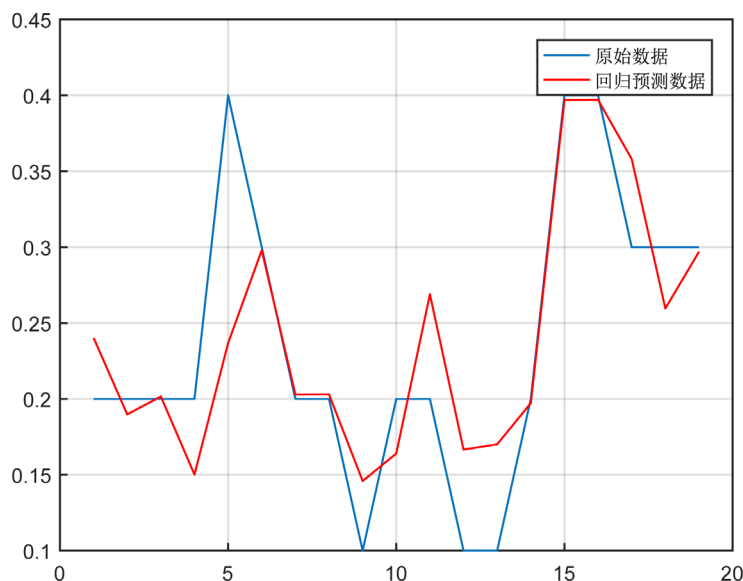


Figure 2. Regression prediction data and the original data comparison chart

图2. 回归预测数据与原始数据对比图

2.2.1. BP 神经网络填补

BP 神经网络是一种具有三层或者三层以上神经元的神经网络，包括输入层、中间层(隐含层)和输出层，上下层之间全连接，而同一层的神经元之间无连接，两个神经元之间的连接强度为网络的权值。BP 算法称为“误差反向传播算法”，通过误差逆向传播修正的反复进行，逐步修正各连接权值，核心是“负梯度下降”理论，误差调整方向沿着误差下降最快的方向进行[5]。

使用表 1 数据为网络学习样本。在本例中将前三个属性作为输入，Petal.Width 作为输出，构成 3 个输入 1 个输出的网络，将前 19 个无缺样本数据作为训练样本集，后 1 个有缺样本作为预测检验样本。

对样本的输入、输出数据进行规格化处理： $\tilde{t} = \frac{2(t-t_{\min})}{t_{\max}-t_{\min}} - 1$ ，其中， $t_{\max}$ 和 $t_{\min}$ 分别为 t 的最大值和最小值， $\tilde{t}$ 为规格化后的变量。MATLAB 中提供了对数据进行规格化处理的函数 $[tn, pn] = \text{mapminmax}(t)$ ，相应的逆处理函数 $t = \text{mapminmax}(\text{'reverse'}, tn, ps)$ ，执行的算法是 $t = 0.5(\tilde{t} + 1) \cdot (t_{\max} - t_{\min}) + t_{\min}$ [6]。

BP 网络用于函数逼近时，权值的调节采用的是负梯度下降法，这种调节权值的方法存在收敛速度慢和局部极小等缺点。

BP 网络的学习过程常发生振动，每一次运行的结果可能会相差很大。如果将隐含神经元个数取为 3，查看某一次运行结果的回归结果图(图 3)和误差直方图(图 4)，可以得知，此时数据拟合的不好，误差也相当大。

对于隐含层的神经元个数的确定，至今尚未找到一个很好的解析式，只能根据经验确定或自行设计一些估计方法[7]。在初始化神经网络时，为了使计算结果相对稳定，往往需要进行多次实验，从而找到使结果最稳定的隐含层神经元个数。

2.2.2. RBF 神经网络填补

RBF 神经网络是模拟视网膜的感受功能产生的：距离感受视野中心越近的视神经元越兴奋；视神经元的激活区域呈径向对称。将视网膜感受原理映射到 RBF 神经网络，可对隐含层神经元进行建模：

$\psi_i(x) = G\left(\frac{\|x - c_i\|}{\sigma_i}\right)$ 。其中， x 为输入样本； c_i 为感受视野中心(中心点)，决定了径向基函数围绕中心点

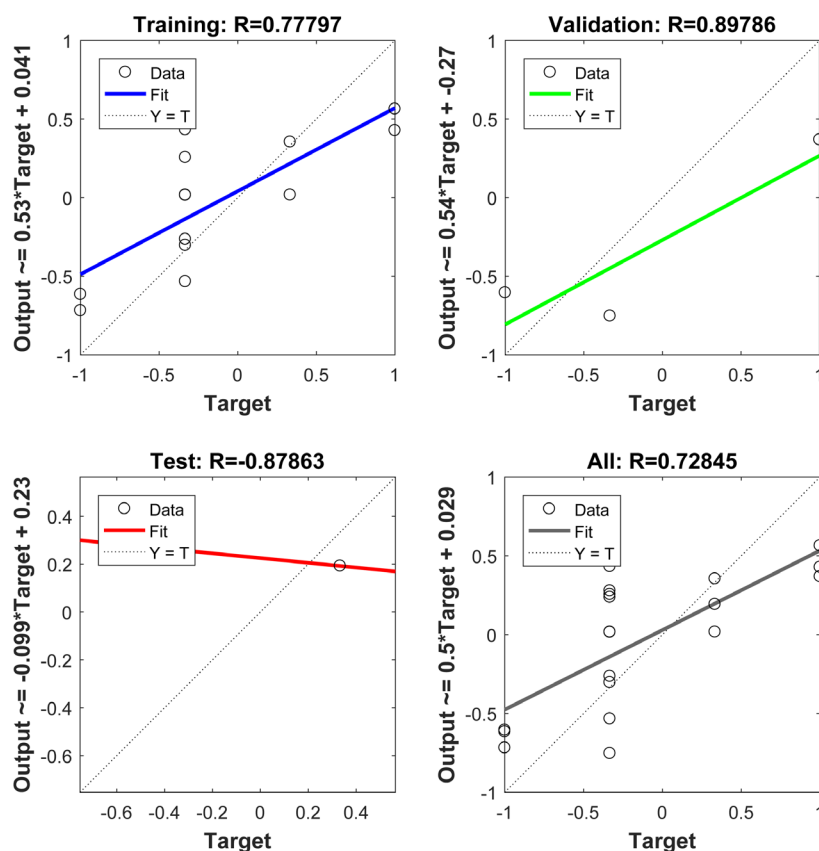


Figure 3. Regression results

图 3. 回归结果图

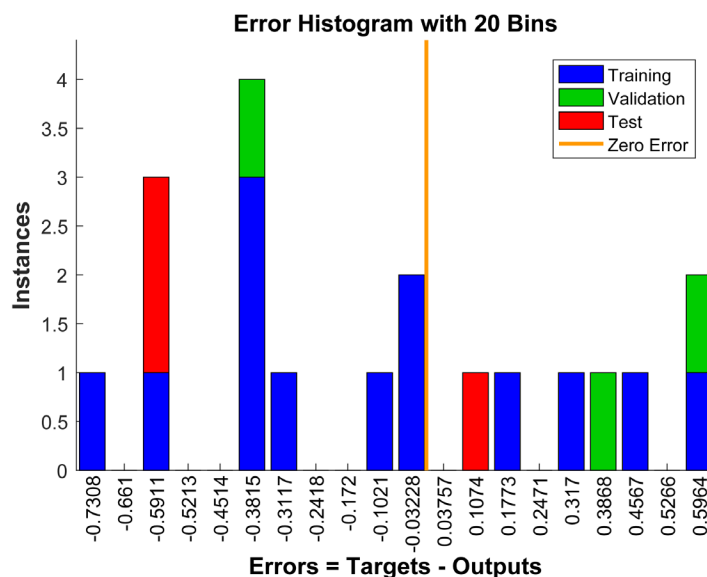


Figure 4. Error histogram

图 4. 误差直方图

的宽度； σ 为宽度； $G(\bullet)$ 为径向基函数，也称激励函数、传递函数或激活函数； $\|x - c_i\|$ 是距离函数，表示网络样本值与数据中心(中心点)之间的距离； $\psi_i(x)$ 为网络输出。隐含层神经元径向基函数通常选用 Gaussian 函数，则上述神经元相应模型转化成： $\psi_i(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma_i^2}\right)$ [4]。

RBF 网络也有三层，隐含层激活函数是高斯函数，第一层(输入神经元)和第三层(输出神经元)一般采用线性激励函数；依据链式偏微分法则，可得网络数据中心、宽度和权值的调整量[8]。利用 MATLAB 提供的神经网络工具箱[9]实现人工神经网络的功能十分方便。RBF 与 BP 的不同点在于，中间隐含层神经元的个数，BP 网络需要根据经验取定，RBF 网络会在训练过程中自适应地取定。依然使用表 1 数据作为网络学习样本，RBF 网络预测第 20 个样本的缺失属性值(Petal.Width)为 0.2489，相对误差为 17.02%。可见，RBF 网络模型的预测结果优于 BP 网络模型的预测结果。

3. 填补缺失的样本类别

选取 iris 数据集中 setosa、versicolor 的样本各 10 个，组成一个新的数据集。删去新数据集中 10 号和 20 号样本的类别，模拟类别值缺失的情况，如表 2。令 setosa、versicolor 为“1”类、“2”类，那么缺失的 10 号和 20 号样本的类别真实值分别为“1”和“2”。

用 $i=1,2,\dots,20$ 分别表示 20 个样本，第 i 个样本的第 j 个属性取值为 a_{ij} 。类别已知的 18 个样本点的均值向量 $\mu = [\mu_1, \mu_2, \mu_3, \mu_4]$ ，标准差向量 $\sigma = [\sigma_1, \sigma_2, \sigma_3, \sigma_4]$ 。对所有样本点数据进行标准化处理：

$\tilde{a}_{ij} = \frac{a_{ij} - \mu_j}{\sigma_j}, i=1,\dots,20; j=1,\dots,4$ 。对应地，称 $\tilde{x}_j = \frac{x_j - \mu_j}{s_j}, j=1,\dots,4$ 为标准化指标向量。记

$\tilde{x} = [\tilde{x}_1, \dots, \tilde{x}_4]^T$ 。记标准化后的 18 个类别已知的样本点数据行向量为 $b_i = [a_{i1}, \dots, a_{i8}], i=1,\dots,18$ 。利用线性内核函数的支持向量机模型进行分类，求得支持向量： b_2, b_9, b_{17} ，线性分类函数： $c(\tilde{x}) = \sum_i \beta_i K(b_i, \tilde{x}) + b$ ，

其中 $\beta_i = \alpha_i g_i$ ， $\tilde{x} = [\tilde{x}_1, \dots, \tilde{x}_4]$ ， $K(b_i, \tilde{x}) = (b_i \cdot \tilde{x})$ 。当 $c(\tilde{x}) \geq 0$ 时， $\tilde{x}$ 属于“1”类；当 $c(\tilde{x}) \leq 0$ 时， $\tilde{x}$ 属于“2”类[6]。

Table 2. A set of biological samples with missing category values
表 2. 有缺失类别值的生物样本集

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	
61	7	3.2	4.7	1.4	versicolor
62	6.4	3.2	4.5	1.5	versicolor
63	6.9	3.1	4.9	1.5	versicolor
64	5.5	2.3	4	1.3	versicolor
65	6.5	2.8	4.6	1.5	versicolor
66	5.7	2.8	4.5	1.3	versicolor
67	6.3	3.3	4.7	1.6	versicolor
68	4.9	2.4	3.3	1	versicolor
69	6.6	2.9	4.6	1.3	versicolor
70	5.2	2.7	3.9	1.4	

注：10 号、70 号样本缺失的类别真实值分别为 setosa、versicolor。

用判别函数判别，得到缺失的第十个、第二十个样本类别分别为“1”、“2”，与真实值一致。所有已知样本点回代分类函数皆正确，误判率为 0。

4. 缺失数据填补效果比较与分析

因为神经网络基于传统统计学，研究样本无穷大时的渐进理论[10]，故以此推导出的各种算法很难在样本数据有限时取得理想的应用效果。同时，因为神经网络是基于经验的风险最小化，不能保证网络的泛化能力[11]，故需要设计者有效利用自己的经验，网络系统的优劣是因人而异的。

与神经网络相比，支持向量机以统计学理论为基础，主要针对小样本情况[12]，能够基于有限的样本信息求解。同时，支持向量机基于结构风险最小化原则，有严格的理论和数学基础[13]，可以避免神经网络实现中的经验成分。

致 谢

大学生创新创业项目(X201610022132)对全基因组关联分析中数据缺失研究。

参考文献 (References)

[1] 陈爱军. 最小二乘支持向量机及其在工业过程建模中的应用[D]: [博士学位论文]. 杭州: 浙江大学, 2006.

-
- [2] MATLAB 中文论坛. MATLAB 神经网络 30 个案例分析[M]. 北京: 北京航空航天大学出版社, 2010: 133-140.
- [3] 王洪波. 单分类支持向量机的学习方法研究[D]: [博士学位论文]. 杭州: 浙江大学, 2012.
- [4] 束洪春, 邬乾晋, 张广斌, 孙士云, 刘可真. 基于神经网络的单端行波故障测距方法[J]. 中国电机工程学报, 2011, 31(4): 85-92.
- [5] 卓金武. MATLAB 在数学建模中的应用[M]. 第 2 版. 北京: 北京航空航天大学出版社, 2014: 99-125.
- [6] 司守奎, 孙兆亮. 数学建模算法与应用[M]. 第 2 版. 北京: 国防工业出版社, 2015: 350-356, 761-776.
- [7] 周永权. 前向多层代数神经网络隐层神经元数目估计方法[J]. 广西民族大学学报(自然科学版), 2000, 6(3): 190-192.
- [8] 卢金娜. 基于优化算法的径向基神经网络模型的改进及应用[D]: [博士学位论文]. 太原: 中北大学, 2015.
- [9] 徐国志, 李茂元. 基于 Matlab 神经网络工具箱进行港口集装箱运量预测[J]. 港工技术, 2003(4): 15-17.
- [10] 张昕. 基于 SVM 方法的医学图像分类研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2006.
- [11] 李巍. 基于宏观经济指标和人工智能方法的上证综合指数预测[D]: [硕士学位论文]. 成都: 西南财经大学, 2012.
- [12] 平源. 基于支持向量机的聚类及文本分类研究[D]: [博士学位论文]. 北京: 北京邮电大学, 2012.
- [13] 刘斌. 支持向量机及其在信号处理中的应用[D]: [硕士学位论文]. 大庆: 大庆石油学院, 2006.

期刊投稿者将享受如下服务:

1. 投稿前咨询服务 (QQ、微信、邮箱皆可)
2. 为您匹配最合适的期刊
3. 24 小时以内解答您的所有疑问
4. 友好的在线投稿界面
5. 专业的同行评审
6. 知网检索
7. 全网络覆盖式推广您的研究

投稿请点击: <http://www.hanspub.org/Submission.aspx>

期刊邮箱: aam@hanspub.org