

一种基于信息熵的数据预处理技术

阮祥超, 万定生

(河海大学 计算机与信息学院, 江苏 南京 211100)

摘 要: 神经网络是数据挖掘领域中的重要研究课题, 针对水文数据不规则和非线性的数据特征, 提出了一种基于信息熵的数据预处理方法, 使用了基于改进的 BP 神经网络模型. 该模型采用计算范围熵的熵值的方法, 遍历查找替换输入变量, 使得输入数据集的变量间线性关系减弱, 非线性关系增加, 熵值增大, 达到提高水文预测精度的目的. 在实际的预测中基于相关信息熵的改进神经网络模型的预测效果优于传统预测模型.

关键词: 数据挖掘; 神经网络; 相关信息熵; 水文预测

中图分类号: TP3

文献标识码: A

文章编号: 1000-7180(2018)02-0104-05

A Data Preprocessing Technique Based on Information Entropy

RUAN Xiang-chao, WAN Ding-sheng

(College of Computer and Information, HoHai University, Nanjing 211100, China)

Abstract: Neural network is an important research topic in the field of data mining. This paper proposing a data preprocessing method based on information entropy which is directed to the irregular and non-linear data characteristics of hydrological data, and uses an improved model of BP neural network. The model uses the method of calculating the entropy of the range entropy, traversing the input variables, searching and replacing them, making the linear relationship between the variables of the input data set weakened. By this way, the non-linear relationship increasing, and the value of the information entropy increasing, so that it can achieving the purpose of improving the accuracy of hydrological prediction. In the actual prediction, the improved model of neural based on the relevant information entropy is superior to the traditional prediction model.

Key words: data mining; neural network; information entropy; hydrological forecasting

1 引言

水利资源的有效管理和应用使得人们不断地挖掘大自然水文现象的规律, 从而有效控制水能源, 预防水灾的发生. 人工神经网络技术作为计算机时代的产物, 不断地被研究者应用于水文预测之中^[1].

神经网络技术作为人工智能学科主要研究内容之一, 对应的模型方法被研究和应用, 但是相同模型对应不同数据集的预测效果却各有各的不同, 如何有针对性地研究输入数据集, 发掘高效的神经网络模型, 需要我们深入研究和探讨^[2]. 在数据挖掘中, 输入变量的选取对预测结果的精准度有很大的影

响^[3-4]. 如果输入变量之间相关性过低则很难找到一个合适的模型来拟合数据, 而训练数据观测项之间太过相似则容易使得模型陷入平坦, 训练效率低下. 因此为了找到更好更适当的输入变量, 并尽可能充分发挥改进算法精确预测的功能, 本文提出了一种基于信息熵的数据预处理方法, 并对 BP 神经网络进行了改进, 相比于传统的 BP 神经网络, 其精确性得到了很大的提高^[5-7].

2 理论基础

2.1 信息熵原理

微观意义上, 熵是系统构成中存在的大量微观

粒子无序程度的量度,系统越有序、越协调,熵就越小;反之系统越无序、越混乱,熵就越大.熵有如下特性:

- (1)熵均大于等于零.
- (2)设 N 是系统 S 内的事件总数,则熵为每个事件概率与概率对数的乘积的和,当事件概率相等时系统 S 的熵最大.
- (3)联合熵: $H(X,Y) = H(X) + H(Y)$,当且仅当 X,Y 在统计学上相互独立时等号成立.
- (4)条件熵: $H(X|Y) = H(X,Y) - H(Y) / H(X)$,当且仅当 X,Y 在统计学上相互独立时等号成立.

2.2 相关信息熵

根据香农信息熵理论,如果随机变量 X 取值为 $X_k(k = 0,1,2,\cdots, n)$ 时的概率为 P_k ,当概率和为 1 时, X 的信息熵 H 定义为:

$$H = \sum -P_k * \log P_k \tag{1}$$

利用信息熵可以定义相关信息熵,设样本为 $n * m$ 矩阵:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} & x_{1(m+1)} \\ \vdots & \ddots & & \vdots & \vdots \\ \vdots & & x_{ij} & \vdots & \vdots \\ \vdots & & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} & x_{n(m+1)} \end{bmatrix},$$

$x_{i(m+1)}$ 为修正后的误差项,则误差项与系统间的相关信息熵 H 定义如下:

$$H' = \sum_{0 < i < n} - \frac{x_{i(m+1)}}{\sum_{0 < p < n} \sum_{0 < q < (m+1)} x_{pq}} \log \left[\frac{x_{i(m+1)}}{\sum_{0 < p < n} \sum_{0 < q < (m+1)} x_{pq}} \right] \tag{2}$$

3 范围熵

对于一个输入数据集,每一个数据点都随机地分布在数据矩阵中.为了描述数据集内部一定范围内的数据统计特征,引入范围信息熵的概念.对于数据集中每一个 $N * M$ 的数据块,都可以视为一个独立的单元.每个单元的信息熵定义为范围信息熵,将第 (i,j) 个单元记为 $\{F(x_i,y_j); i = 1,2,3\cdots, j = 1,2,3\cdots\}$,该数据单元中数据出现概率定义为

$$P_{ij}(L) = \sum \sum F(x_i,y_j) / AREA_{ij} \tag{3}$$

$$AREA_{ij} = \sum F(x,y) \tag{4}$$

对于一个复杂数据集,可以通过对整个数据集分块进行循环遍历后,得到各个数据块中数据出

现的概率和范围熵.随着数据块中数字的递增和数据块颜色的变浅,逐层逐块完成整个数据集的遍历,最终得到某个数据在数据集中出现的总概率 $P(x,y)$,如图 1 所示.根据得到的各个数据块的范围熵可以确定一个最佳阈值,这个阈值就是该数据集的最佳分割值.一个最佳阈值可以将一个完整的数据集合分割成 A,B,C,D... 等等不同的子部分,如图 2 所示.

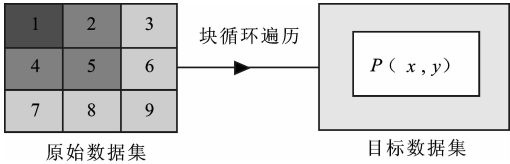


图 1 块循环数据集示意图

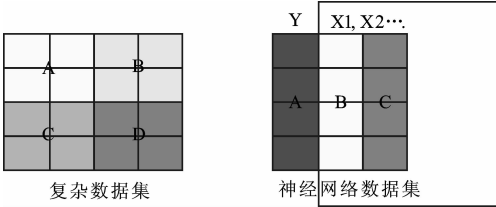


图 2 不同类型数据熵模型

这些子数据块也可以称作数据块熵等高线分割,不同分割的子数据块类的数据具有明显的差异性.在数据挖掘中,这些差异性往往会导致分析的奇异性.

对于一个确定的输入数据集而言,我们关心的往往不是数据集内部的等高线分割,关注的是目标列 Y 和输入列 X 之间的差异.换言之,我们关注的不是数据块之间的熵差异,而是数据列之间的熵差异.复杂数据集的熵模型和本文使用的数据集的熵模型可以如图 2 所示.

从物理意义上讲,数据挖掘的需要分析的数据集合中,一行数据即为一个观测.不同的观测往往是针对同一个事物在不同条件,时间序列下的观测值,因此,行之间必然存在某种联系,而列之间则可以视为是输入变量之间的关系.这种关系直接反映了各种因素对分析结果的影响,因此在数据挖掘中我们往往将范围熵定义为

$$\hat{P}_{ij}(L) = \sum F(x_i,y) / AREA_{ij} \tag{5}$$

$$Y = \sum y \tag{6}$$

$$AREA_{ij} = \sum F(x,y) \tag{7}$$

4 基于范围熵的数据处理算法

熵改进的核心思想是,通过熵改进算法去除数据集中对建模有影响的冗余信息和奇异数据.冗余信息意味着需要分析的数据集中有着重复的数据或者数据模式,某些数据的属性是多余的和不必要的.冗余信息是一种过剩的信息,它不仅仅是不必要的,而且对数据分析是有害的.因为冗余数据常常导致过度训练,对冗余数据进行训练往往会扰乱数据分析的结果.但是合适的冗余可以减少错误的概率.奇异信息即是数据集中由于种种原因而出错的数据点.这种错误可能在数据输入时,甚至是原始数据获取时就产生了.在数据挖掘中称这些点为孤立的数据点,它们的存在会极大的影响模型的分析与训练.

提高系统的相关信息熵一般有两种方法.

方法 1:使用合适的编码函数对数据集进行编码.

数据编码的目的是使得输入数据集适合分析方法,通过数据编码可建立数据间的内在联系,将原本散乱的数据统一成一定的格式,这样就间接地提高了系统的熵值.常见的数据编码有曼彻斯特编码、循环编码、卷积码、哈弗曼编码等.

方法 2:调整数据集使之有序.

调整数据集的目的是弱化数据集之间的线性关系,使得系统熵值提高.一般的需要分析数据集从物理意义上而言,一行数据即为一个观测,不同的观测往往是针对同一个事物不同条件、不同时间下的观测值,因此,行之间必然存在某种联系.而列之间则可以视为是输入变量之间的关系.这种关系直接反映了各种因素对结果的影响,因此我们通过对某一观测列与另一观测列的相对差值后,可以使得系统线性关系弱化,即使得系统的熵值有所提高.

设定变量间熵的最佳阈值,根据基于范围熵的数据处理算法,选取原始数据集中提取的某列作为参考列,并将该列保存在缓存中.逐个计算参考列与其他数据列之间的范围熵,经过遍历计算后,超过最佳阈值的数据列将被相关信息熵值最大的改进数据列替换.经过熵的遍历查找和新变量替换后,得到的有序数据集中新变量间线性关系减弱,非线性关系增强.熵改进的过程如图 3 所示.

算法分为熵改进过程和 BP 神经网络建模过程两个步骤.算法的流程图如图 4 所示.

算法首先计算列模块间的相互信息熵即范围熵,如果相互信息熵的分析结果不满足要求则调用

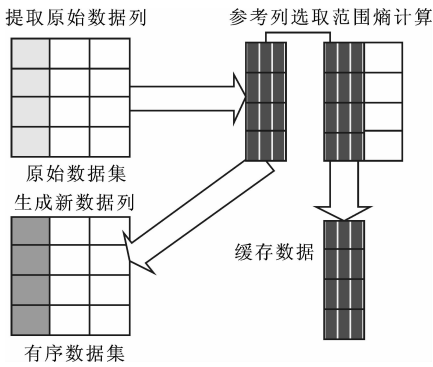


图 3 熵改进过程

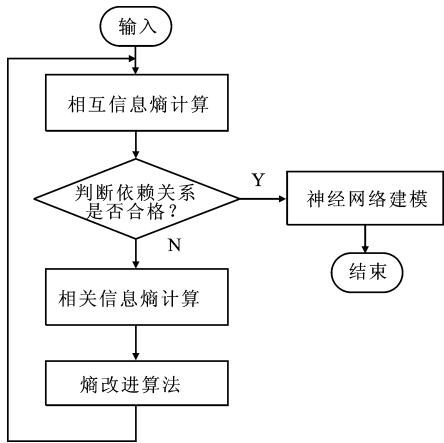


图 4 基于范围熵的数据处理算法流程图

熵改进算法对数据集进行调整.该算法通过迭代运算改进数据矩阵的信息熵.熵改进算法如算法 1 所示.

算法 1 熵改进算法.

INPUT:初始数据集合

OUTPUT:调整后的数据集合

- (1)BEGIN:输入初始数据集矩阵,指定需要计算相互信息熵的项
- (2)计算相互信息熵是否达到要求,如果没有达到要求则 go to 3
- (3)计算相关信息熵 A
- (4)temp=输入数据集;
- (5)BEGIN x=i;j
- (6)tempA=temp[x]
- (7)tempB=temp[x+1]
- (8)data [x]=(tempB-tempA)
- (9)END;
- (10)根据式(7)求出更新后的相关信息熵 B
- (11)对比 A 和 B 相关信息熵是否有改进.是,则 go to 3;否,则令 data=temp
- (12)return data
- (13) END;

5 应用研究

5.1 数据准备

本文的实验数据选自南京市下关、大通、江阴、镇江四个站点的汛期的潮(水)位数据作为输入数据进行模型的训练和测试. 其中,选择下关站 1999~2006 年 1252 条数据作为训练集(training set),选择 2006~2007 共 277 条数据作为测试集(test set). 经过数据预处理方法获得输入列的初始输入数据格式(见表 1),其中行表示时间轴,列表示数据所在的站点. 第一列实测值代表输出比较列,从第二列向后的几列代表数据输入列,作为原始的输入数据带入模型的训练之中.

表 1 原始数据格式								
xgT +1	dtT -1	dt T	xg T	jyT -2	jyT -1	jy T	zjT -1	zj T
5.63	8.86	8.97	5.48	4.81	4.72	4.47	5.43	5.19
5.67	8.97	9.18	5.63	4.72	4.47	4.32	5.19	5.04
5.72	9.18	9.43	5.67	4.47	4.32	4.21	5.04	4.98
5.82	9.43	9.67	5.72	4.32	4.21	4.06	4.98	4.90

5.2 案例分析

设定变量间熵的最佳阈值,根据基于信息熵的改进神经网络算法,针对某列变量做变量之间的范围熵的熵值计算,超过最佳阈值的,找到相关信息熵值最大的改进数据列替换该列. 经过熵的遍历查找和新变量替换后,变量间线性关系减弱,非线性关系增加,变量与变量间的熵值增大(熵越大,系统越无序),系统的熵提高.

在改进过程中,输出列下关明日真实值数据不参与熵值的计算中,只针对输入变量做范围熵改进变量替换,因此预测结果能够更真实有效的反映下关的预测趋势. 熵改进模型对预测误差精度提高的区分统计结果显示如表 2 所示.

表 2 熵改进模型预测结果区分统计表		
区分统计标准名称	区分统计标准符号	熵改进模型
训练均方差	MSE1	0.015 292
测试均方差	MSE2	0.021 187
拟合部分平均绝对误差	TR_MAE	0.069 459
区分绝对误差	TR_MSE	0.010 79
测试部分平均绝对误差	TS_MAE	0.089 458
测试区区分绝对误差	TS_MSE	0.014 949

从图 5 中可以看出,经过熵模型改进后,下关水位明日和今日的差值相对缩小,实际误差在区间

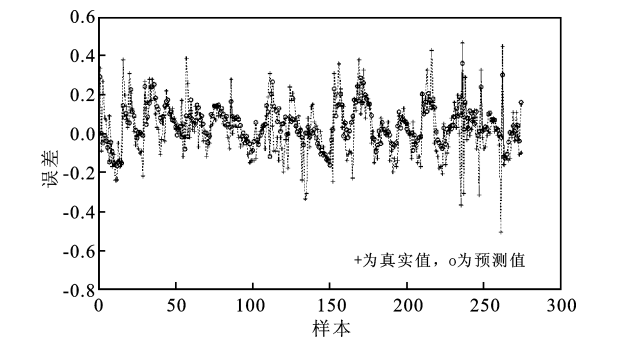


图 5 预测变量误差精度变化

$[-0.6, 0.6]$ 之间波动,预测误差在区间 $[-0.2, 0.4]$ 之间波动,表明预测误差精度得到了提高.“o”曲线代表经过模型改进后预测误差项的数值分布,“+”曲线代表实测误差项的数值分布.

改进模型通过输入数据集的范围熵计算,实现了对输入变量的选择和更新数据集的目的. 从预测结果表 3 可以看出,预测误差数据还原后的下关明日预测值有效的实现了下关明日数据的预测.

表 3 预测值误差还原数据样例				
实测误差	预测误差	下关今日值	下关明日值	下关明日预测值
-0.15	-0.147 7	4.87	4.96	4.722 3
-0.74	-0.046 2	4.96	4.92	4.913 8
-0.98	-0.093 4	4.92	4.75	4.826 6
-1.41	-0.155 1	4.75	4.64	4.594 9

熵改进神经网络模型通过范围熵的概念,提出了根据熵值计算的方法替换原始输入变量. 图 6 显示出,熵改进的方法从预测数据的误差精度上得到了明显的提高,使得神经网络强大的处理非线性数据的功能得以充分的发挥.

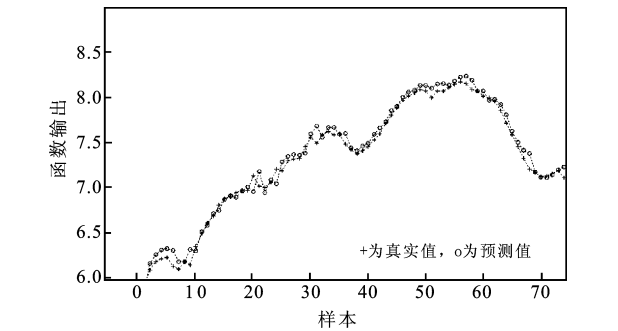


图 6 熵改进神经网络算法预测图

6 结束语

本文提出一种基于信息熵的数据预处理技术,采用改进的 BP 神经网络模型对下关水位数据进行

预测. 通过这样一种技术,减弱了新加入的误差项与原来的变量之间相关性,降低了数据集的冗余性和奇异性,减少了变量间的线性特征,从而提高了系统的信息熵,增加了神经网络效能,使得预测结果更加准确.

参考文献:

[1] 张继学,王鹏,张琳,等. 人工神经网络在短期降水预测方面的应用研究[J]. 科技风, 2016, 28(17): 123-124.

[2] 王智文. 基于改进 BP 神经网络的车牌字符识别[J]. 微电子学与计算机, 2011, 28(9):66-69.

[3] 高武奇,康凤举,钟联炯. 数据挖掘的流程改进和模型应用[J]. 微电子学与计算机, 2011, 28(7):9-12.

[4] 潘笑,万敏. 基于模糊神经网络的数据挖掘方法研究[J]. 微电子学与计算机, 2005, 22(12):48-50.

[5] Wu B, Han S, Xiao J, et al. Error compensation based on BP neural network for airborne laser ranging

[J]. Optik - International Journal for Light and Electron Optics, 2016, 127(8):4083-4088.

[6] Jia H, Pan D, Yuan Y, et al. Using a BP Neural Network for Rapid Assessment of Populations with Difficulties Accessing Drinking Water Because of Drought [J]. Human & Ecological Risk Assessment, 2015, 21(1):100-116.

[7] Lu C, Ma L, Min Y, et al. Regional information entropy Demons for infrared image nonrigid registration [J]. Optik - International Journal for Light and Electron Optics, 2016, 127(1):227-231.

作者简介:

阮祥超 女,(1994-),硕士研究生. 研究方向为信息处理与信息系统. E-mail: xchruan@hhu. edu. cn.

万定生 男,(1963-),教授. 研究方向为信息处理与信息系统.

(上接第 103 页)

[2] 吴丰,吕振雷,陈珂锐. 基于时间尺度分析的云网格空间资源调度算法[J]. 计算机仿真, 2015, 32(8): 131-135.

[3] 嵇可可. 基于动态趋势预测蚁群算法的云计算资源调度优化研究[J]. 科技通报, 2016, 32(1):187-190.

[4] 邱昕. 一种新型的基于效用函数的 LTE-A 系统预编码资源调度算法[J]. 科学技术与工程, 2015, 15(6): 215-219.

[5] 徐文忠,彭志平,左敬龙. 基于遗传算法的云计算资源调度策略研究[J]. 计算机测量与控制, 2015, 23(5):1653-1656.

[6] 崔永君,张永花. 基于特征尺度均衡的 Linux 系统双阈值任务调度算法[J]. 计算机科学, 2015, 42(6): 181-184.

[7] 曹一家,刘易珠,阙凌燕,等. 换电站与电网协调的多

目标双层实时充放电调度方法[J]. 电力自动化设备, 2015, 35(4):1-7.

[8] 孔凡凤,欧红玉,龙林德,等. 基于连通支配集的 WSN 自适应数据调度算法[J]. 计算机工程, 2015, 41(10):94-98.

[9] 顾炜江. 物联网中共享资源数据调度方法的研究[J]. 计算机仿真, 2017, 34(1):268-271.

[10] 林智华,高文,吴春明,等. 基于离散粒子群算法的数据中心网络流量调度研究[J]. 电子学报, 2016, 4(9):2197-2202.

作者简介:

罗 曼 女,(1986-),硕士,实验师. 研究方为向计算机应用、多媒体应用.

E-mail: ss356398632@163. com.