

# 基于监督学习的数据预测服务构建方法

李 昭,宋 壹,陈 鹏

(三峡大学 计算机与信息学院,湖北 宜昌 443002)

**摘 要:**数据处理、分析、预测是当前计算机行业发展的新增长点,也是经济社会不断进步的网络技术支撑。为更好地从数据中挖掘隐式特征和隐性关系,进一步提高数据预测的命中率、准确性,依托所研发的科研大数据服务平台提出了基于监督学习的数据预测服务构建方法。通过样本采集和特征提取、特征预处理、建模技术选取的步骤建立用于数据预测的数学模型,进而基于服务平台构建数据预测服务,同时结合平台共建共享、操作便捷等优势,提升数据预测服务的实用性和复用性。以新闻延时预测为实验用例,在平台中使用前向逐步线性回归和三维点云建模技术构建预测服务,通过10-折交叉验证对服务性能进行度量。实验结果表明,该方法复用性强,所构建的服务可对数据进行有效预测,为用户进行准确决策提供支持。

**关键词:**监督学习;数据预测;服务构建;前向逐步回归;三维点云;交叉验证;新闻延时

**中图分类号:**TP301

**文献标识码:**A

**文章编号:**1673-629X(2019)09-0188-07

**doi:**10.3969/j.issn.1673-629X.2019.09.036

## Data Prediction Service Construction Based on Supervised Learning

LI Zhao, SONG Yi, CHEN Peng

(School of Computer and Information Technology, China Three Gorges University, Yichang 443002, China)

**Abstract:** Data processing, analysis and prediction are new growth points for the development of computer industry, and also the support of network technology for the continuous progress of economic society. To better mine implicit characteristics and implicit relations from data, and further improve the hit ratio and accuracy of the data prediction, we put forward a data prediction service construction method relying on the existing research big data service platform. The mathematical model of data prediction is established by sample collection, feature extraction, feature preprocessing, modelling technology selection, etc, and then the data prediction service is built based on the service platform. Meanwhile, combined with the advantages of platform co-construction and sharing, convenient operation, etc, the practicability and reusability of data prediction services are improved. Taking the news delay prediction as an experimental case, the prediction service is constructed by using forward progressive linear regression and three-dimensional point cloud modeling technology on the platform, and the model performance is measured by 10-fold cross validation. Experiment shows that the proposed method has strong reusability, and the service built can predict the data effectively, providing support for users to make accurate decisions.

**Key words:** supervised learning; data prediction; service building; forward stepwise regression; three-dimensional point cloud; cross-validation; news delay

## 0 引 言

现代信息产业的长足发展使人们逐渐从对信息数量的崇拜转向对信息质量的追求。作为计算机信息产业、大数据领域的重要组成部分,针对数据预测的研究近年来不断取得新的成果,呈现出了蓬勃发展的势头。随着数据量的不断增长,该领域相关技术不断发展成

熟,这也为数据预测赋予了更加丰富的内涵:从分析角度看,数据预测是对数据信息内在本质、潜在关联的深入挖掘与剖析;从应用角度看,数据预测是对数据信息增长方向、发展趋势的准确评价与预估。数据量的膨胀虽然为数据预测领域提供了广阔的素材空间与研究基础,但另一方面也制约了预测技术的进一步优化,集

收稿日期:2018-10-10

修回日期:2019-02-12

网络出版时间:2019-03-28

**基金项目:**国家重点研发计划项目(2016YFC0802500);国家自然科学基金(61272236);湖北省自然科学基金(2018CFC852);教育部人文社科规划基金(20171304);三峡库区地质灾害教育部重点实验室开放基金(2015KDZ05);三峡大学人才专项基金(8000303)

**作者简介:**李 昭(1986-),男,博士,副教授,CCF会员(27657M),研究方向为数据挖掘、人工智能、业务流程挖掘;宋 壹(1996-),男(土家),硕士研究生,CCF会员(97629G),研究方向为机器学习、软件测试。

**网络出版地址:**<http://kns.cnki.net/kcms/detail/61.1450.TP.20190327.1633.072.html>

中表现为以下几点:一是噪声数据的大量存在扰乱了预测方法的正常工作,使得预测效率降低;二是数据安全受到严峻挑战;三是缺乏有针对性、深层次的信息分析提炼手段,数据的价值未得以充分发挥<sup>[1]</sup>。大数据时代处理数据理念的三大转变是“要全体不要抽样、要效率不要绝对精确、要相关不要因果”<sup>[2]</sup>,因此,掌握好、挖掘好、运用好既有数据,不断从数据中创造更多的价值,成为了数据预测研究领域的新课题。

文中研究内容所依托的“三峡大学科研大数据计算服务平台”是根据数据预测领域研究趋势和发展目的构建的具有鲜明应用导向的开放型服务平台,涵盖了信息上传、内容分析、模型构建等内容,为数据预测服务方法的构建提供了现实可用的载体。该平台的一大亮点是数据互通互用、方法共建共享,抽象包装好的数据上传模块大大提高了上传效率、优化了上传体验,而且数据一旦被上传到云端,所有平台用户均可查看、下载;平台还为用户提供了数据预测服务的构建体系,包括数据访问、数据预处理、特征工程、统计分析、机器学习、文本分析、数据可视化等模块,用户可以根据自己的实际需要,以抽象的方法构建出相应的服务模型,结合自己或其他用户已经上传的数据即可投入实际运行使用。例如,基于数据采集与特征分析的城市火灾风险预测服务、城市人口疾病概率预测服务、影片受欢迎程度预测服务、新闻节目延时风险预测服务等,都可以基于该科研大数据计算服务平台进行构建和实现。

在数据预测服务构建方面,文献[3]只指出了数据预测的理论背景和应用领域,没有对相关方法做进一步研究;文献[4]提供了数据预测模型性能度量的维度与相关技术,但没有通过实际用例进行实验研究;文献[5]提出了基于主成分分析和统计建模的数据预测模型,但仅仅应用在经济预测领域,没有抽象出可移植的通用模型。

作为对该平台实用性、可靠性的验证,文中通过数据样本采集和特征提取、特征预处理、建模技术选择等过程,提出了一种基于监督学习的数据预测服务构建方法。该方法以机器学习中的监督学习为基本手段,构建了“数据—特征—模型—数据”的预测链,较高的抽象性使其移植性能良好,能够在比较广泛的领域得以应用,从而为科研大数据平台上具体服务的构建提供统一化模型。

1 科研大数据服务平台简介

数据量的迅猛增长在为数据使用者带来机遇的同时,也催生了许多亟待解决的问题。有的研究者掌握大量数据,但空白的数据整合方法、落后的数据建模技术、低下的数据使用效率制约了数据量优势的发挥;有

的研究者有一套科学系统的数据分析机制和模型构建体系,但匮乏的数据获取渠道使研究工作缺乏宝贵的原材料。数据与技术不相适应的矛盾已经成为数据预测领域的一个重要瓶颈。

将数据与技术进行有机整合的科研大数据服务平台为解决这一矛盾提供了新路径。该平台的一个重要优势是将“数据上传—数据分析—数据应用”这一封闭管道改造成了开放链条,实现了一人上传、多人分享、群体共用。具体地说,当构建一个数据预测服务时,一个用户将原始数据上传至服务器云端,该数据可以立即被平台上的其他用户检索、浏览到,而且基于该数据的预测模型构建过程也可以由所有用户一起完成,所得到的模型结果可以一起应用,这有效地提高了对数据潜在价值的挖掘能力。

2 数据预测服务的构建方法

用好科研大数据服务平台的关键在于拥有一个好的数据预测服务构建方法。数据预测服务构建方法相当于平台上的一个抽象“模具”,以它为基础可以构建出各种不同的数据预测服务,从而在各个领域有针对性的发挥作用。因此,数据预测服务构建方法对于整个平台能否有效运行具有十分重要的意义:一个好的构建方法可以为各个服务的构建提供良好的模板,从而提高运行效率、减少错误产生的可能性;相反,一个坏的构建方法不单单影响自身,依据它所创建的具体服务都会带有先天缺陷,从而严重影响平台的运行效果。

文中提出的基于监督学习的数据预测服务构建方法(见图1),以样本、特征、建模技术三个对象为主体,具有较好的可扩展性和可移植性。

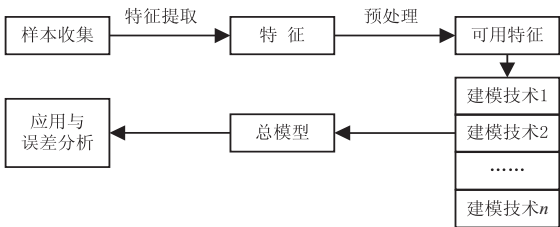


图1 基于监督学习的数据预测服务构建方法

2.1 样本采集与特征选取

数据样本是数据预测的基本对象,也是整个研究过程的开展空间,采集样本是研究开始的前提。在一般的科研过程中,样本采集的方式主要可以归为以下几类:一是在相关领域的信息公开网站上直接下载,如文献[6];二是通过 API 接口在线获取,如文献[7];三是通过人工方式手动采集。无论通过何种方式,采集的数据都必须符合真实、完整、客观、准确的要求,才能够应用到下一步的分析中,以保证实验结果的可靠性。

在监督学习中,采集的数据样本总和称为样本空间,它会被进一步划分为训练集和测试集,以构建模型并进行检验。

特征是对研究对象的高度抽象,是数据对象所含信息的代表性表示,是表征数据的关键。在样本采集过程中,一个数据对象往往包含着大量信息,其中部分与研究相关的信息对工作起到了重要作用,但大部分信息与研究工作无关或关联度较小,将其纳入研究范围会大大降低数据分析、预测的效率。因此,从大量信息中找准提取数据特征的角度,成为了每一名研究者必须面对的问题。选取特征既可以根据生活常识、工作经验进行人工判断,也可以辅助 SVD 分解技术<sup>[8]</sup>,通过计算能量值并设定取舍阈值选取最具影响力的特征。总之,特征的选取一定要符合两个方面的要求:一是最大限度表征数据样本;二是最大程度降低计算开销。

## 2.2 特征预处理

通过相关方法收集到的特征往往不能直接投入后续算法进行应用,这可能与数据本身的特质有关,也可能与待使用算法模型对数据的要求有关。如果在样本采集的过程中出现欠采样或者过采样<sup>[9]</sup>问题,导致样本类别不均衡,则需要增加或减少相应样本;如果采集到了大量的异常样本,而这些样本本身并无太大实际意义并且对模型的构建起到了严重的负面作用,则需要对样本的选择与清洗;如果特征向量中某一维的取值范围过大,而其实际影响力与其他特征并无显著差别,则需要对特征进行归一化处理。

预处理既是对数据样本的进一步提取与精炼,也是对下一步输入模型的准备与铺垫,它并不产生新的对象,只是通过在既有特征对象上施加映射关系,生成一种新的表示。预处理方式的选择一是要为数据预测的最终目的服务,二是要符合特征的本质属性,三是要契合后续待使用模型的相关要求。

## 2.3 建模技术的选择

根据处理好的特征进行建模是数据预测服务构建方法的最后一步,也是最重要的一步。在一般的建模过程中,往往只对特征选取一种技术进行建模,这在预测要求较为简单的情况下应用得比较广泛,但在特征数量丰富、特征间关系复杂的情况下则不再适用。文中描述的数据预测服务构建方法提出在同一数据集上分别使用不同建模技术,以提高总模型与数据的拟合程度,进一步优化预测效果。

在该方法中,服务是平台的实例,模型是服务的载体,因此,选择好的建模技术对单个服务乃至整个平台的质量具有决定性意义。需求导向是对建模技术进行选择的根本遵循,即构建的服务需要产生什么样的结

果,就相应地选择什么样的模型;同时也要考虑模型与数据特征的兼容性,确保模型不仅能用得好,还能用得稳。

## 3 对构建方法的实例验证

本部分用一个具体的应用服务来验证以上提到的构建方法。

### 3.1 服务应用背景概述

近年来,随着新闻舆论工作的全面加强,新闻数量的不断扩大、新闻内容的不断增多,导致电视新闻节目的既定时长经常无法满足实际的播出需要,延时<sup>[10]</sup>情况频频出现。特别是中央电视台《新闻联播》节目,延时频率、幅度呈现出了“双上升”势头。以全国“两会”召开的3月为例,2016、2017、2018年3月《新闻联播》节目延时的次数分别为6次、12次、22次,月延时率同比分别上涨了19.1%和32.3%;另据统计,在2017年9月下旬至2018年9月下旬的365期《新闻联播》中,延时节目期数为75,延时率高达20.5%,相当于每五天就有一次延时情况发生。

频繁出现的延时情况会对电视台生产播出各环节造成连锁影响:一是打破节目常规播出预案,播出线上的各种不确定因素显著增多,播出事故风险陡然上升;二是影响后续节目编排,尤其是《新闻联播》之后的黄金时段节目,会因延时出现播出时间后移、节目时长缩减甚至取消播出等严重后果;三是广告播出受到波及,每天19点30分之后的广告具有数量少、价格高、影响大、传播广等特点,每秒钟均价高达数万元,延时使得广告无法按时播出造成经济损失;四是地方卫视也会因此受到影响,国家有关部门明确规定地方台每晚需完整转播央视《新闻联播》节目,延时情况的出现会使所有地方台不得不临时做出调整。

从《新闻联播》大量的历史播出库中提取分析相关数据,对可能出现的延时情况进行定量研究成为了预测延时、减小风险、降低损失的新途径。

### 3.2 样本采集和特征选取

《新闻联播》的延时具有一定的时间聚集性,在一些重大事件发生的时间段,延时的几率高于平时。虽然近年来该节目的延时次数大幅增加,但相较于每天播出一期的密度,延时率依然维持在较低区间,“不延时是常态,延时是例外”的基本面没有打破。为有效分析《新闻联播》延时特点,准确找出延时背后的关键因素,选取了较具延时代表性的2015年9月、2016年3月和10月、2017年3月、2017年9月中上旬的135期节目;同时考虑到更为普遍的一般性,选取了2017年9月下旬至2018年9月下旬的365期节目,组成容量为500的样本空间。



特征是对样本的概貌性描述,是表征样本的关键点,抓好特征是用好样本的基础与前提。在新闻延时预测服务中,时政新闻字数、占比及辐射指数三个特征可以较好地对样本进行解释。特征间关系如图2所示

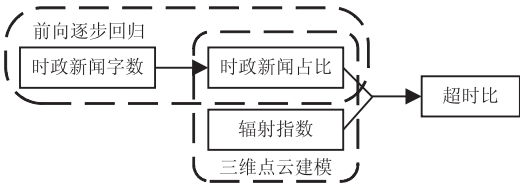


图2 特征间关系

3.2.1 时政新闻字数

时政新闻一般指党和国家的重要会议、国家重大外事活动及领导人出席的活动等。不同于其他类型新闻,时政新闻是《新闻联播》的必播内容,可变性小、播出弹性小,用其衡量节目的延时风险具有较好的代表性。随着新兴媒体的不断发展,绝大多数时政新闻的文字稿件在每晚《新闻联播》播出之前就会被官方媒体发布于网站,因此时政新闻字数(word number, WN)可以作为特征用于预测。

3.2.2 时政新闻占比

时政新闻占比(current politics ratio, CPR)定义为时政新闻时长(current politics duration, CPD)占节目常规时长(1 800 秒)的比例,它是预测延时风险的一个重要指标,如式1所示。

CPR =  $\frac{\text{CPD}}{1\,800}$

(1)

例如,当该特征值在0.5时,说明时政新闻时长为15分钟,余下15分钟可用于其他类型新闻的播放,延时风险较低;当该特征值在0.8时,留给其他类型新闻的播放时间仅剩6分钟,延时风险较高;当该特征值在1及以上时,说明仅时政新闻就已达到或超过30分钟,延时风险为100%。

《新闻联播》播音员语速近年来处于较为固定的区间范围,所以WN和CPD之间存在着增长关系,该关系可以利用前向逐步线性回归方法找到。而由式1可知,CPD与CPR之间呈现出线性关系,所以可以由WN直接得到CPR。利用前向逐步线性回归算法找到这一关系的过程将在3.4节具体描述。

3.2.3 辐射指数

单靠时政新闻占比预测节目的延时风险有时并不可靠。当CPR很高时,节目可能会压缩或者取消排序靠后的社会新闻、国际新闻,以对冲延时风险;当CPR很低时,也有可能大量播放与时政新闻配套的其他新闻,从而造成超时比(overtime ratio, OR,实际播出时长与节目常规时长的比值)升高。

在抽取的500个样本中,延时样本有107个,其平

均CPR为0.78,但其中也有部分样本CPR值非常小;非延时样本有393个,其平均CPR为0.28,其中也有部分样本CPR值非常高。表1列出了部分此类异常样本。

表1 CPR与OR不相适应的部分异常样本

CPR	OR	CPR	OR
0	1.267	0.866	1
0.041	1.667	0.823	1
0.103	1.333	0.774	1
0.112	1.167	0.758	1
0.189	1.5	0.719	1

这类异常样本出现的原因在于忽略了《新闻联播》节目编排中的要闻影响因素。当处于重大活动及节日期间时,《新闻联播》节目为配合活动的开展、营造节日的氛围,会有意地增加相关新闻的播出量,而这往往对是否延时及延时幅度造成较大影响。为此,文中提出“辐射指数”(influence exponential, IE)特征,表征重大活动及节日对《新闻联播》节目延时的影响程度,如式2所示。

$$IE = \sum_{i=1}^k \frac{\alpha_i \beta_i}{|distance_i| + 1}$$

(2)

其中,k表示某天附近范围内可能对当天新闻节目时长产生影响的重大事件数;distance为该事件与当天的时间距离;α为该事件影响力大小的量化体现;β(初始值置0)根据该事件所处时间位置表示其是否对当天节目产生影响,如产生则赋值为1,否则保持初始值。

根据对往期《新闻联播》节目播出规律的观察,提炼出党代会开闭幕、全国两会开闭幕、重大外交活动和其他重要活动等四类对延时率贡献较大的主要事件,其α、β及影响邻域取值由表2定义。

表2 辐射指数公式相关参数取值规则

事件类型	α	影响邻域	β
党代会开闭幕	3	[-3,+3]	1
全国两会开幕	3	[-3,+7]	1
全国两会闭幕	3	[-7,+3]	1
重大外交活动	2	[-2,+2]	1
其他重要活动	1	[-1,+1]	1

如2018年全国两会的开幕时间分别是3月3日、3月5日,闭幕时间分别是3月15日、3月20日,根据定义的影响邻域及相关指数,可以划出这四个事件在当月的影响范围,如图3所示。

以3月10日为例,其处于事件1、事件2、事件3三个事件的影响半径内,因此有:

$$IE = \sum_{i=1}^3 \frac{\alpha_i \beta_i}{|distance_i| + 1} = 1.38$$

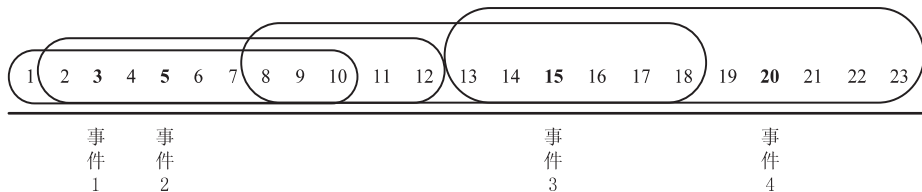


图 3 2018 年 3 月重大活动影响范围

### 3.3 特征预处理

在挑选出来的特征中,CPR、IE、OR 均为个位数,而 WN 则多以千、万为单位,这给特征间相互关系的挖掘带来了负面影响。为此,首先对 WN 进行归一化处理,将其转化为分布于 0-1 之间的值。

为使有限的数据集发挥出更好的效能,有必要对数据集进行合理划分。这里采用基于分层采样<sup>[11]</sup>的 10-折交叉验证<sup>[12]</sup>,将 500 个样本均分为 10 个子集,每个子集的非延时样本与延时样本之比控制在 4 : 1 左右。

### 3.4 建模技术选择一:通过前向逐步回归预测 CPR

时政新闻占比是预测延时比的重要指标,但每天的 CPR 只有当节目播出后才能获得,因此单纯的 CPR 对预测没有直接意义。3.2.2 中已经提到可以通过机器学习算法,用时政新闻字数来预测时政新闻占比,从而将 CPR 这一后得特征转化为先得特征,达到预测的目的。

#### 3.4.1 前向逐步回归算法

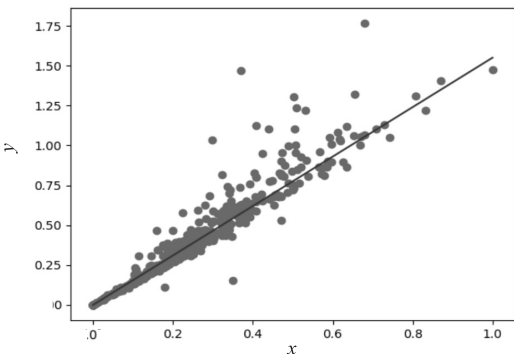
前向逐步回归是机器学习线性回归算法<sup>[13]</sup>中的一个重要方法,其将误差初始化为无穷大,之后对特征赋予初始值为 0 的权重,通过每次对权重加、减步长后计算并覆盖误差,得到使误差最小的系数。(它属于一种贪心算法,每一步都尽可能减少误差<sup>[7]</sup>)

在此,采用绝对值误差度量真实值与预测值之间的偏差,如式(3)所示。

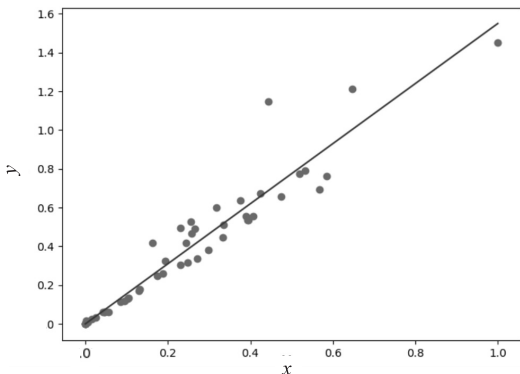
$$RSS = \sum_{i=1}^k (|y_i - f(x)_i|) \tag{3}$$

#### 3.4.2 建立 WN-CSR 模型

这里展示了以第一个子集为测试集,后九个子集为训练集得到的 WN-CSR 前向逐步回归模型,图 4 (a)、(b)分别是将该模型应用在训练集和测试集上的



(a) 将模型应用在训练集



(b) 将该模型应用在测试集

图 4 前向逐步回归建模过程及模型拟合效果结果(  $x$ 、 $y$  轴分别表示归一化处理后的 WN 和 CSR )。

可以看到,该模型对大多数样本点进行了很好的拟合,但仍有部分样本点与模型相距较远,且实际值高于预测值的“正向误差样本量”远多于实际值低于预测值的“反向误差样本量”,即时政新闻字数较少时仍有较高的几率出现高时政新闻占比。文献[14]对此现象给出了解释。

尽管如此,大多数样本间依然存在较为明显的线性关系,经过 10-折交叉验证,可得 WN-CSR 平均模型为:

$$y = 1.539\ 2x \tag{4}$$

该模型平均训练误差、平均测试误差分别为 0.05、0.14。

### 3.5 建模技术选择二:通过三维点云建模预测 OR

3.2.3 节给出了计算辐射指数 IE 的公式,3.4.2 节给出了 WN-CSR 模型,本节讨论利用 MATLAB 的 cftool 工具箱进行三维点云建模,得到 CSR-IE:OR 的映射关系。

#### 3.5.1 Curve Fitting Tool

MATLAB 提供了大量实用的工具箱,其中 cftool (curve fitting tool) 因其“使用方便、功能强大、能实现多种类型的线性或非线性曲线”<sup>[15]</sup>而得以广泛应用。它包含了多种对数据点进行逼近和拟合的方式,在建模完成后还会提供拟合度、自由度、均方误差等指标,为用户判断该模型的好坏提供量化依据;友好的数据可视化功能也是该工具箱的一大亮点。cftool 为三维点云曲面拟合提供了四种方式,即 custom equation (自定义方程)、interpolant (插值逼近)、LOWESS (局部加权回归散点平滑) 及 polynomial (多项式拟合)。在较

为常用的多项式拟合中,需要用户指定较为合适的最高幂次;如果幂次过高,模型对数据学习得太好而泛化能力较差,就会出现“过拟合”<sup>[16]</sup>;幂次过低可能导致模型无法挖掘到数据间的内在关系,从而不能充分逼近数据,即出现“欠拟合”现象。cftool 工具箱在曲面拟合时允许的最高幂次为“双 5 次”<sup>[17]</sup>。此外,cftool 也为用户提供了指定模型鲁棒性的机会,可根据实际需要选择 off(常规最小二乘法)、LAR(最小绝对值残差)和 bisquare(二次方权值)<sup>[18]</sup>。

3.5.2 建立 CSR-IE;OR 模型

该模型的构建过程仍然采用 10-折交叉验证方式,将第 1~10 个子集依次作为测试集,其余 9 个子集

依次作为训练集。在当前训练集、测试集上,将三维点拆分出  $X$ 、 $Y$ 、 $Z$  轴作为 cftool 的输入。在输入参数选项中,拟合方式选择多项式拟合“Polynomial”,幂次选择“ $x:2,y:2$ ”,鲁棒性选择最小绝对值残差“LAR”。

由此可得该模型的一般形式,如式 5 所示。

$$f(x,y)=p_{00}+p_{10}x+p_{01}y+p_{20}x^2+p_{11}xy+p_{02}y^2$$

(5)

其中, $x$  为时政新闻占比 CSR; $y$  为辐射指数 IE; $f(x,y)$  为超时比 OR。

图 5 展示以第一个子集为测试集,后九个子集为训练集得到的 CSR-IE;OR 三维点云模型( $x$ 、 $y$ 、 $z$  轴分别表示 CSR、IE、OR)。

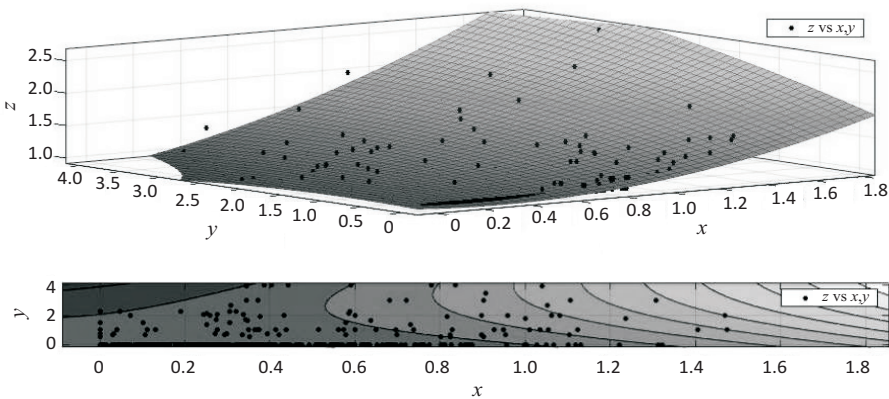


图 5 CSR-IE;OR 点云模型

经过 10-折交叉验证可得 CSR-IE;OR 平均模型:  
 $p_{00}=1.006\ 6, p_{10}=-0.142\ 6, p_{01}=0.057\ 3, p_{20}=0.330\ 7, p_{11}=0.173\ 1, p_{02}=-0.030\ 4$ 。

以上步骤给出了该方法构建出的一个具体数据预测服务,该服务在科研大数据服务平台上的部署如图 6 所示。



图 6 服务方法在数据平台上的部署

4 模型应用与误差分析

3.5.2 节给出了 10 次实验得出的平均模型。为进一步评估模型的可信度,从训练误差、测试误差、决

定系数(R-Square)、均方根误差(RMSE)、可信度、延时可信度、非延时可信度等七个维度对 10 个模型进行度量。

训练误差指该模型在对应训练集上的平均误差;测试误差指该模型在对应测试集上的平均误差;决定系数<sup>[19]</sup>取值范围为 $[0,1]$ ,表征模型对数据的解释能力,越接近 1 表示拟合程度越高;均方根误差<sup>[20]</sup>表征预测值与实际值的离散程度(见式 6);可信度为模型在每一组测试集 50 个样本中预测成功的比例;延时可信度为模型在每一组测试集所有延时样本中的查出率;非延时可信度为模型在每一组测试集所有非延时样本中的查出率。具体如表 3 所示。

$$R-Square = \frac{RSS}{TSS} = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

(6)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

(7)

依据表 3 可知,该方法的平均训练误差、测试误差均在 0.06 左右,即预测新闻节目时长的误差均值为 1.8 分钟;平均决定系数为 0.981 6,说明预测模型与真实模型的拟合度处于较高水平;均方根误差为 0.027 5,说明预测值与真实值离散程度较小;可信度

为 0.86,说明该方法判断新闻节目延时与否的可信度为 86%;延时可信度为 70.4%,说明该方法判断新闻节目延时的可信度为 70.4%;非延时可信度为 0.911,

说明该方法判断新闻节目不延时的可信度为 91.1%。综上所述,基于监督学习的数据预测服务构建方法在新闻延时领域取得了良好的应用效果。

表 3 模型误差度量结果

实验次数	训练误差	测试误差	决定系数	均方根误差	可信度	延时可信度	非延时可信度
1	0.060 9	0.071 9	0.982	0.027 1	0.92	0.769	0.973
2	0.066	0.045	0.982 2	0.027 4	0.86	0.667	0.902
3	0.070 4	0.050 5	0.981 8	0.028 2	0.8	0.5	0.875
4	0.066 4	0.103 9	0.982 4	0.027	0.76	0.563	0.912
5	0.063 7	0.040 6	0.982 3	0.027 7	0.94	0.9	0.95
6	0.060 7	0.069 1	0.978	0.027 9	0.92	0.889	0.927
7	0.066 2	0.078 7	0.982 1	0.027 3	0.82	0.615	0.892
8	0.062	0.069 4	0.982 4	0.027 4	0.84	0.6	0.9
9	0.063 3	0.073 6	0.981 6	0.027 9	0.86	0.714	0.884
10	0.060 8	0.064 1	0.981 1	0.027 2	0.88	0.818	0.897
平均	0.065 3	0.067 8	0.981 6	0.027 5	0.86	0.704	0.911

5 结束语

文中提出的基于监督学习的数据预测服务构建方法,以科研大数据服务平台为依托,以实际应用中的不同需求为导向,以具体服务为实际的运行载体,对整个服务构建过程提供了一套流程完善、可用性和复用性强的机制。以新闻节目延时预测为例进行的实验表明,所构建的服务对数据进行了合理采集,对特征进行了准确抽取,对建模技术进行了有效选择,最终获取了良好的预测结果。

一个好的数据预测服务构建方法既需具备良好的实用性和复用性,也需最大程度实现用户的预测需求。文中提出的基于监督学习的数据预测服务构建方法在实用性和复用性上表现良好,但方法可变性不足,仍需在特定的应用领域进行优化。

参考文献:

[1] 陈 光. 基于大数据的数据服务应用研究[J]. 计算机技术与发展,2018,28(8):129-134.

[2] 孔 钦,叶长青,孙 赟. 大数据下数据预处理方法研究[J]. 计算机技术与发展,2018,28(5):1-4.

[3] 刘 婧,姜文波,邵 野. 计算机视觉艺术在数字媒体中的应用[J]. 信息与电脑:理论版,2018(11):164-165.

[4] 李尚晋. 大数据环境下的机器学习研究[J]. 电子世界,2018(1):62-63.

[5] 王 丽,李 阳,蓝 尉,等. 基于主成分分析和统计建模的数据预测[J]. 工业控制计算机,2018,31(7):123-124.

[6] 卢月明,王 亮,仇阿根,等. 局部加权线性回归模型的PM2.5空间插值方法[J]. 测绘科学. 2018,43(11):79-84.

[7] HARRINGTON P. 机器学习实战[M]. 北京:人民邮电出版社,2013:213-215.

[8] 洪 泓,邓志新,张大山,等. 基于SVD振动提取算法的视

频麦克风[J]. 工业控制计算机,2018,31(7):74-75.

[9] 季晨雨. 不平衡数据分类问题解决办法[J]. 电子技术与软件工程,2018(15):152-153.

[10] 艾 达. 央视《新闻联播》10年的变与不变[J]. 新闻与写作,2013(3):22-25.

[11] 汪海涛,余永奎,段春雨. 基于大数据不平衡样本集的重采样方法及应用[J]. 现代计算机,2018(22):26-29.

[12] 周志华. 机器学习[M]. 北京:清华大学出版社,2016:26-27.

[13] JAIN M B,NIGAM M K,TIWARI P C. Curve fitting and regression line method based seasonal short term load forecasting[C]//2012 world congress on information and communication technologie. Trivandrum,India:IEEE,2012:332-337.

[14] 贾曜榕. 浅谈新闻联播播音员的语速变化[J]. 戏剧之家,2015(7):127.

[15] 胡尊乐,纪小敏,闫 浩,等. 基于 Cftool 拟合工具箱的中田舍河水位流量关系暨产汇流模型的构建[J]. 江苏水利,2018(6):1-7.

[16] CHANDRASHEKAR G,SAHIN F. A survey on feature selection methods [J]. Computers and Electrical Engineering,2014,40(1):16-28.

[17] 黄兵锋,解方喜,傅佳宏. MATLAB 曲线拟合工具箱在发动机特性拟合中的应用[J]. 湖北文理学院学报,2014,35(5):26-28.

[18] 尹宏俊,邢思茗,乔月俊. Matlab 软件在费用模型建立中的应用[C]//中国设备管理协会寿命周期费用委员会第七次年会论文集. 武汉:中国设备管理协会,2011:151-156.

[19] ISRAELI O. A Shapley-based decomposition of the R-Square of a linear regression[J]. The Journal of Economic Inequality,2007,5(2):199-212.

[20] KALBIL S,FALLAH A,BETTINGER P,et al. Mixed-effects modeling for tree height prediction models of Oriental beech in the Hyrcanian forests[J]. Journal of Forestry Research,2018,29(5):1195-1204.