doi:10.3772/j.issn.1002-0470.2019.01.003

基于多输出神经网络的舆情分析指标拟合及优化研究①

陈 娟②* 王功明③** 徐翼龙*** 王海威****

(*北京大学新闻传播学院 北京 100190)
(**中国科学院生物物理研究所 北京 100101)
(***北京联合大学智慧城市学院 北京 100101)
(****军委后勤保障部信息中心 北京 100842)

摘 要 通过互联网媒介数据构建出完整的互联网舆情指标体系,是进行舆情预测及评估、网络空间治理的基础。然而,由于数据冲突、数据不完整、计算误差、标注失误等诸多问题,严重降低某些指标的可信度。本文根据可信度高低将舆情指标划分为两类,综合多变量数据拟合、主成分分析(PCA)、多输出神经网络等技术,以及基于数据类型的指标评价方法,能够由高可信度指标推导出低可信度指标,并采用新浪微博用户数据进行性别判断实验与用户粉丝量实验。实验结果表明,所推导出的性别准确率高达96.7%,用户粉丝量的相对绝对误差(RAE)为16%,说明本方法可以构建高可信度舆情指标体系,为舆情指标体系的构建和量化研究奠定基础。

关键词 舆情指标体系,可信度,指标拟合,主成分分析,多输出神经网络

0 引 言

指标体系是由若干相对独立且彼此联系、反映社会经济现象总体特征的统计指标所形成的树形分类结构。指标体系的建立是预测或评价社会经济现象的前提和基础,它是将抽象研究对象按照其本质属性和特征的某一方面分解成为具有行为化、可操作化的结构,并对指标体系中每一构成元素(即指标)赋予相应权重的过程[1]。舆情是指民众对于社会事件或社会现象发表的意见、观点、情绪和态度的总和及其传播,在影响公众视听、引导社会意识形态及政府公共事务决策方面发挥越来越大的作用[2]。要科学地掌握舆情的发展趋势并采取合适行动,就需要构建舆情指标体系。构建舆情指标体系旨在将舆情信息定量化,有助于全面了解舆情的发生、发展与趋势,及时通过预警帮助管理者掌握舆情处置的主动权[3]。

目前,舆情指标体系研究主要是从不同角度建立指标体系,包括:舆情发展变化特征[4]、舆情传播组成要素及产生根源[5]、舆情专题属性[6]、舆情演变规律[7]、舆情涨落的内外作用力[8]等。无论采用什么方式建立舆情指标体系,关键问题是获取高精度指标元素。在获取过程中,不同指标的获取难度存在差异,有些指标(如用户必填信息)可以直接获取,可信度较高;有些指标(如用户选填信息)存在不完整、不一致等缺陷,有些指标(如最大值、最小值、均值、方差等统计特征)是二次加工的结果,有些指标(如用户类别、情绪等)是人工标注的结果,这些指标的可信度相对较低。

虽然不同指标的可信度存在差异,但都处于反映特定社会经济现象的同一指标下,因此具有较高的相关性,可以建立不同可信度指标间的拟合关系,

① 国家自然科学基金(61502475,61841601)资助项目。
② 女,1991 年生,硕士生;研究方向:舆情大数据;E-mail: chenjuan@ict.ac.cn
③ 通信作者,E-mail: gongmingwang@126.com
(收稿日期:2018-08-06)

验证已有指标的准确性,提高已有指标的可信度,也可以预测难以获取的指标,丰富和完备指标体系。赵建伟[9]采用替换法优化指标中的无效值和空值,但效果受到周围数据的影响,没有用到其他有效指标的信息。孙新伟[10]采用贝叶斯分类器建立性别、用户忠诚度、用户品类群体等指标的预测模型,但仅适用于生成用户标签,应用范围有限。姚龙飞[11]通过云模型和本体思想构建模型,预测用户画像标签,但应用范围亦有限。冯娟娟等[12]将画像标签划分为事实标签、模型标签、预测标签 3 种类型,并综合多种分类/聚类算法、推荐算法、机器学习方法进行预测,但上述标签都是由指标体系的基础数据推导出来,如果基础数据的可信度不好,就会影响标签的预测准确性。曾鸿和吴苏倪[13]根据关联规则联系用户行为信息和偏好,提高用户需求预测的准确性,但所用的关联规则存在诸多限制性前提,普适性差。鉴于此,本文综合主成分分析(principal component analysis,PCA)、多输出神经网络等方法,建立根据不同可信度指标转换模型,能够根据已有高可信度指标推导出低可信度指标,辅助建立完整的舆情指标体系。

1 关键技术

1.1 多变量数据拟合

在数据分析中,需要围绕反映待分析对象,根据所涉及的标准、规则、算法、资源等要素,采集不同种类的数据,这些数据变量构成特定场景下的指标体系。在采集过程中,不同变量的获取难度存在差异,第一类可以直接获取,从而具有较高的可信度;但是,第二类变量存在不完整不一致等缺陷、或是在直接测量基础上二次加工的产物、或是人工标注的结果,这类变量的可信度相对较低。在这种情况下,需要提高第二类变量的可信度。

虽然这两类变量的可信度存在差异,但都处在同一个指标体系中,彼此之间的相关性应该要高于指标体系外的其他变量[14]。所以,通过对可信度较高的变量进行数据拟合,从理论上可以推导出那些可信度较低的变量。

进行多变量数据拟合,主要包括两个步骤,即主成分分析、多变量预测。其中,主成分分析对可信度较高的变量进行降维,减少计算复杂度,常用方法是主成分分析(PCA);多变量预测可以根据降维后变量推算出那些可信度较低的变量,常用方法是多输出神经网络。

1.2 主成分分析

主成分分析也称主分量分析,是常用的数据降维方法。其本质是一种数学变换方法,它把给定的一组相关变量通过线性变换转成另一组不相关的变量,这些新的变量按照方差依次递减的顺序排列,称为主成分。在实际应用中,根据特征值和累计贡献率,确定所选用主成分的数量;因此,主成分的数量小于原有变量的数量,但和原有变量存在较强的相关关系,可以视为原有变量的线性组合。在这种情况下,对主成分进行分析,其效果和直接分析原有变量近似等效,但主成分数量远小于原有变量,所以实现了较好的降维效果[15]。

主成分分析法的基本流程如下:(1)对原始变量按列排列并标准化,(2)计算标准化后矩阵的相关系数矩阵,(3)计算相关系数矩阵的特征值和特征向量,(4)根据特征值和累计贡献率确定主成分个数,(5)计算主成分实现降维。

1.3 多输出神经网络

神经网络是模仿动物神经网络行为特征,通过内部大量结点相互关联,并行执行信息处理的数学模型。最基本的神经网络是反向传播(back propagation, BP)神经网络(BPNN),通常由输入层、隐含层、输出层组成[16]。常规 BP 神经网络只有 1 个隐含层,在大多数文献中隐含层均小于 3 层,其原因是隐含层增加后权系数数量会显著增加,给计算和存储带来沉重负担,传播误差也逐渐增大。

近年来出现的卷积神经网络(convolutional neural networks, CNN)可以解决上述问题,数据在进入全连接隐含层之前,预先经过输入层、卷积层、激励层、池化层的处理,分别执行预处理、局部卷积滤波、非线性映射、池化压缩等操作,数据质量得到很大程度的提高。

但是,CNN 的输入元个数固定,在某些输入数

据数量不稳定的场景,如智能问答、机器翻译,就无法投入使用;此外,CNN 没有记忆功能,无法处理时间序列类型数据。在这种情况下,产生了递归神经网络(recurrent neural, RNN)。

与 CNN 相比,RNN 隐藏层之间的结点存在连接,隐藏层既可以接收当前时间点输入层传递的信息,也能够接收上一时间点隐藏层传递的信息,RNN 会对历史信息进行记忆,并将历史信息应用到当前神经元的输出计算中。RNN 中的神经元具有自反性,是处理序列和列表数据的最佳神经网络模式。

在网络结构上,CNN 每层参数不相同,但是 RNN 每层参数共享复用,即 RNN 中除了输入数据之外,每步信息传递机制均相同。这种特性降低了训练 RNN 时的参数数量,减轻了网络学习的工作量,很大程度上减轻了模型的训练难度。

但是,RNN 实际可用的信息时间跨度有限。对于距离当前时刻时间较长的历史信息,RNN 的学习能力会逐渐减弱,其原因在于梯度回传效应随时间跨度增加而快速衰减,无法保留较长时间段的信息。为了解决这个问题,出现了长短时记忆神经网络(long shortterm memory, LSTM)。

LSTM 是建立在 RNN 上的一种新型深度机器学习神经网络,在输入、反馈与防止梯度爆发之间建立一个长时间的时滞,强制记忆单元的内部状态保持持续误差;可以弥补 RNN 的梯度消失和梯度爆炸、长期记忆能力不足等问题,能够真正有效地利用长距离的时序信息。

1.4　模型评价

按照数据类型,待拟合指标可以划分为 2 类:数值型和类别型。前者是具体的数值,例如注册用户数、销售额等;后者是分类的级别,例如性别、职业、学历等。拟合这两类指标时,需要选择不同的评价方法。

1.4.1　数值型指标评价方法

假设待评估样本个数为 n, 第 i 个样本($1 \leqslant i \leqslant n$)的真实值、预测值分别为 $X_{\text{model},i}$、$X_{bos,i}$, 则均方根误差(root mean square error, RMSE)如式(1)所示。

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{bos,i} - X_{\text{model},i})^2}{n}} \tag{1}$$

相对绝对误差(relative absluteerror, RAE)如式(2)所示[17]。

$$RAE = \frac{\sum_{i=1}^{n} | X_{bos,i} - X_{\text{model},i} |}{\sum_{i=1}^{n} | X_{\text{model},i} - X_{average} |} \tag{2}$$

1.4.2　类别型指标评价方法

假设分类数为 T, 用 TP_v 表示被正确分到第 v 类的元素个数,用 FN_v 表示属于第 v 类但被误分到其他类的元素个数,用 FP_v 表示被误分到第 v 类的元素个数,则准确率和召回率分别如式(3)和式(4)所示[18]:

$$P = \frac{\sum_{v=1}^{T} TP_v}{\sum_{v=1}^{T}(TP_v + FP_v)} \tag{3}$$

$$R = \frac{\sum_{v=1}^{T} TP_v}{\sum_{v=1}^{T}(TP_v + FN_v)} \tag{4}$$

F 值是精确率和召回率的加权调和平均值,表达对精确率/召回率的不同偏好,如式(5)所示。

$$F = \frac{(\alpha^2 + 1) \times P \times R}{\alpha^2 \times (P + R)} \tag{5}$$

在式(5)中, P 和 R 分别是精确率和召回率, α 是调和因子,一般情况下, $\alpha = 1$。

1.4.3　多模型 K 折交叉验证

由 1.3 节可以看出,多输出神经网络有多种类型,包括 BPNN、CNN、RNN、LSTM 等,在类型相同的情况下,通过修改网络结构,又可以得到不同的模型。对于固定的应用场景,通常选择不同的模型进行训练,然后从中筛选效果最佳的模型。

为了保证评估结果的公平性,划分训练集和测试集时尽量保证数据分布一致性。采用 K 折交叉验证选择最合适的模型[19]。

设样本集 S 包含 m 个样本,可供选择的 t 个模型是 M_1、M_2、……、M_t, 进行 K 折交叉验证,具体流程如下。

步骤 1　随机将样本集 S 划分成 k 个不相交的子集,每个子集中样本数量为 m/k, 这些子集分别记作 S_1、S_2、……、S_k。

步骤 2　对每个模型 M_j, $j = 1,2,\cdots,t$, 进行如

下操作。

For $r = 1$ to k

{

将 $S_1 \cup \cdots \cup S_{r-1} \cup S_{r+1} \cup \cdots \cup S_k$ 作为训练集;

训练模型 M_j, 得到相应的假设函数 $H_{j,r}$;

将 S_r 作为验证集,计算模型 M_j 的泛化误差 $\varepsilon_{S_n}(H_{j,r})$;

}

计算 $\varepsilon_{S_n}(H_{j,r})$, $r = 1,2,\cdots,k$ 的平均值,得到模型 M_j 的平均泛化误差。

步骤 3 计算所有模型的平均泛化误差,从中选择平均泛化误差最小的模型 M_{opt}, 该模型就是筛选出来的最佳模型。

一般地,对于数值型指标拟合模型,使用平均绝对百分比误差表示泛化误差;对于类别型指标拟合模型,使用 F 值表示泛化误差。

2 算法描述

2.1 基本流程

算法的处理过程如图 1 所示。

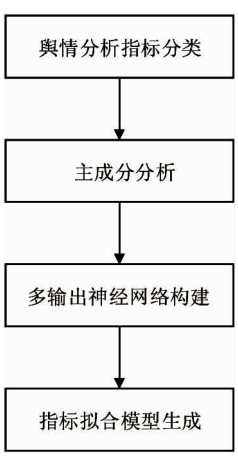

图 1 算法处理过程

算法的整个执行流程包括 4 个阶段。首先,根据舆情指标将数据集划分为两类:容易获取且可信度高的数据集 X、不易获取或可信度差的数据集 Y, 并根据取值情况将 Y 划分为数值型数据集 Y_v 和类别型数据集 Y_c; 然后,通过主成分分析对 X 降维,转换为规模较小的数据集 Z; 随后,分别对组合数据集 (Z, Y_v) 和 (Z, Y_c) 进行训练,得到若干数值型/类别型指标拟合模型;最后,通过 K 折交叉验证,得到满足要求的指标拟合模型。

2.2 舆情分析指标分类

设完整的舆情分析指标体系包含 h 个指标,按照获取难易程度和可信度,这些指标可以划分为 2 类。

(1)容易获取且可信度高的指标。从单个信息源采集且普遍存在的数据,例如:用户姓名、文章数量、登陆日期、点赞次数等。这类指标构成数据集 $X = [x_1, x_2, \cdots, x_p]$, $p \leqslant h$。

(2)不易获取或可信度差的指标。包括在某些信息源缺失的数据(例如,用户信息中的非必填项,不同用户填写的完整性也不相同);在多个信息源存在但不一致的数据(例如,用户在不同网站注册的住址不同,同一数据的差异性会严重影响其可信度);对多个信息源进行汇聚计算后的数据(例如,用户发表的所有文章数量),由于计算误差会影响其可信度;需要人工标注的指标(例如,事件类型、用户情绪等),由于带有标注人员的主观因素,可信度相对较差。这类指标构成数据集 $Y = [y_1, y_2, \cdots, y_q]$, $q = h - p$。

根据指标的取值类型,将 Y 划分为 2 部分:第 1 部分为数值型数据集 $Y_v = [y_{v1}, y_{v2}, \cdots, y_{vg}]$; 第 2 部分为类别型数据集 $Y_c = [y_{c1}, y_{c2}, \cdots, y_{cs}]$, 满足 $Y = Y_v \cup Y_c$, $\boldsymbol{\Phi} = Y_v \cap Y_c$。

完整的舆情分析指标体系包括上述所有指标,任何一种指标的缺失、不一致、计算或标注失误,都会影响指标体系的应用效果。因此,需要提高第 2 类舆情指标的质量。

2.3 指标降维

设样本个数为 m, 则数据集 X 构成规模为 $m \times p$ 的矩阵 $(m > p)$。对该矩阵进行标准化后,计算相关系数矩阵的特征值和特征向量;根据给定的累计贡献率 R_{ac} (一般在 60% 以上,多数情况下是 85%),按照特征值大小选择对应的特征向量,并将特征向量作为列向量,构成转换矩阵 $\boldsymbol{U}$。设所选择的特征向量个数为 l, 则 $\boldsymbol{U}$ 是规模为 $p \times l$ 的矩阵。

执行矩阵乘法 $\boldsymbol{Z} = \boldsymbol{X} \times \boldsymbol{U}$, 可以将数据集 X 转

换为主成分 $\boldsymbol{Z}$,其规模是 $m \times l$, 由于 $l < p$, 所以实现了降维。

2.4　多输出神经网络构建

对于数值型指标,根据 $\boldsymbol{Z}$ 和 Y_v 建立多输出神经网络,输入层和输出层的神经元个数分别是 l 和 g,中间处理层可以选择 BPNN、CNN、RNN、LSTM 等多种类型,由此得到多种不同形态的多输出神经网络 NPV_1、NPV_2、…、NPV_w, 都是适用于数值型指标的拟合模型。

同理,对于类别型指标,根据 Z 和 Y_c 建立多输出神经网络 NPC_1、NPC_2、…、NPC_f。

2.5　指标拟合模型生成

对于数值型指标,采用 K 折交叉验证分析所有构建的多输出神经网络模型,选择泛化误差最小的模型,作为数值型指标拟合模型,记做 NPV。同理得到类别型指标拟合模型 NPC。

因此,该指标体系的拟合模型是由 NPV 和 NPC 构成的混合模型,如式(6)所示。

$$NP = \begin{cases} NPV & \text{数值型指标} \\ NPC & \text{类别型指标} \end{cases} \tag{6}$$

3　实验分析

3.1　数据介绍

本实验使用西安交通大学计算机科学系 MOEKLINNS 实验室搜集到的新浪微博用户数据。其中包括 750 名用户的账户信息与他们发表过的文本数据,总计4.2万条,数据介绍如表1所示。此外,本文使用北京师范大学中文信息处理研究所与中国人民大学 DBIIR 实验室的研究者共同开源的预训练词向量表示文本数据的分布特征。

表 1　实验数据集

目标	属性	属性描述
作者性别	用户名	用户名的文本
	用户简介	用户简介的文本
	用户发表的微博	用户发表的微博文本
用户粉丝量数量	用户等级	用户在微博平台的账户等级
	微博数量	用户之前发表过的微博文章总数
	正负面情绪	用户之前发表过的正面情绪文章比例
	转发量	用户之前发表过的文章被转发的数量

3.2　模型训练

所选择数据对应指标体系包含的指标数量众多,选择全部指标进行验证耗费极大,也不切合实际。为了更好地体现评价方法的效果,本文分别选择指标“作者性别”和“用户粉丝量”作为类别型指标和数值型指标的代表,制定 2 个实验,即性别判断实验与用户粉丝量实验,并分别使用类别型指标评价方法与数值型指标评价方法进行评价。

3.2.1　性别判断实验设计

性别判断实验将作者性别作为预测目标,将用户名、用户简介与用户发表的微博作为属性数据,使用 RNN 与 CNN 神经网络进行学习与预测。

在预处理过程中,首先采用 jieba 中文分词工具对中文文本进行分词[20],并去除文本中的数字与标点符号,并将它们索引化表示;然后综合索引化表示结果与预训练词向量的前200 000个词语,完成深度学习的特征表示。实验中,使用的分词长度均为文本分词结果长度的平均值加减 2 倍方差。文本数据长度,如表 2 所示。

表 2　文本数据长度

参数	分词结果的平均长度(个)	分词结果的最长长度(个)
用户名	2.3	12
用户简介	13.6	56
微博文本	15.9	99

本实验采用 keras 搭建深度学习框架,底层使用 TensorFlow,编程语言使用 Python ;主要运行环境包括 PyCharm 软件、Win10 系统、8 GB内存等。本实验模型主要分为 3 部分:第 1 部分为一层 RNN 或 CNN,用于处理用户名、用户简介、用户发表的微博;第 2 部分为融合层,连接第 1 部分的 3 个网络模型;第 3 部分为 sigmoid 分类器。其他实验参数,如表 3、表 4 所示。

在实验过程中,首先分别使用门控循环单元(gated recurrent unit, GRU)、CNN 与 LSTM 模型对

用户名、用户简介与用户发表的微博进行学习;然后分别使用 5 折交叉验证法对这 3 种算法进行学习,并对比它们的准确率、精确率、召回率与 F 值(调和因子 $\alpha = 1$);最后使用网络模型学习训练集占总体数据集为 20%、40%、60%、80% 时的准确率与 $F1$ 值(调和因子 $\alpha = 1$),进行对比。

表 3　性别判断实验模型参数

参数	值
词向量维度	300
隐含层节点	300
Loss 函数	binary_crossentropy
Optimizer	Adam
Merge 层	concatenate
迭代次数	10

表 4　用户粉丝实验模型参数

参数	值
第 1 层隐含层个数	4
第 2 层隐含层个数	3
第 3 层隐含层个数	2
Loss 函数	MSE
Optimizer	sgd
迭代次数	100000

3.2.2　用户粉丝量实验设计

本实验将用户的粉丝量作为预测目标,将用户等级、微博数量、正负面情绪、转发量作为属性数据,使用 BP 神经网络完成学习与预测。通过分析用户粉丝量的分布范围,发现存在极端数据,为了消除极端数据对模型训练的影响,本实验将粉丝数平均值加 3 倍方差作为上限,并将粉丝数1 000作为下限,用于剔除极端数据。

本实验采用的框架、编程语言与运行环境和性别判断实验完全一致。本实验模型包括 4 类:不包含隐含层的 BP 神经网络、包含 1 层隐含层的 BP 神经网络、包含 2 层神经网络的 BP 神经网络。

实验中分别使用 5 折交叉验证法对这 3 种算法进行学习,并对比它们的 RMSE 与 RAE。

3.3　模型分析

3.3.1　用户性别判断实验结果分析

(1) 使用 5 折交叉验证法,所计算出的准确率(ACC)、精确率(P)、召回率(R)与 $F1$ 值,如表 5 所示。

表 5　性别判断算法准确度比较

模型	ACC	P	R	$F1$
GRU	0.9670	0.9493	0.9622	0.9557
CNN	0.9250	0.8821	0.9208	0.9007
LSTM	0.9560	0.9335	0.9486	0.9410

从表 5 的实验结果可以看出,这些方法在性别判断领域的准确度从大到小依次为 GRU > LSTM > CNN,即 GRU 模型优于 LSTM 与 CNN 模型。

根据精确率与召回率的大小顺序可以看出 GRU 模型效果最好。但是二者在 CNN 与 LSTM 的数值表现上存在歧义,其原因是精确率与召回率在特定情况下彼此矛盾。

而 F 值可以看作精确率和召回率的加权调和平均。当调和因子 $\alpha = 1$ 时,$F1$ 值从大到小依次为 GRU > LSTM > CNN,且基本比例与准确度一致,可以很好地表现 3 种模型的优劣。

(2) 3 个模型的准确率与训练数据集的比例关系,如图 2 所示。

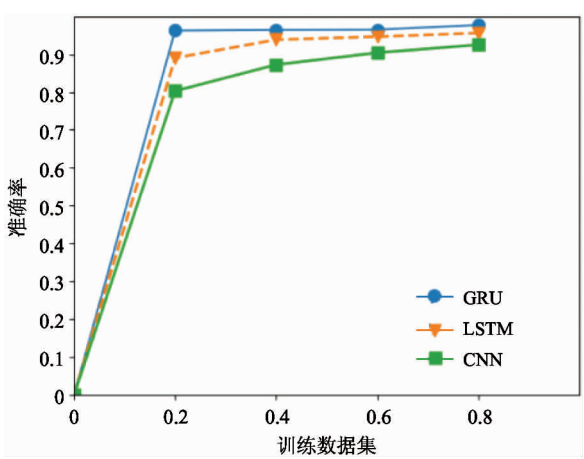

图 2　3 个模型准确率的变化情况

3 个模型的 $F1$ 值与训练数据集的比例关系,如图 3 所示。

在图 2 与图 3 中,圆形表示 GRU 对应的值,三角形表示 LSTM 对应的值,方形表示 CNN 对应的

值。

从图2和3可以看出,准确率与*F*1值都随着训练数据集规模的增加而逐步提高,但是*F*1值能够更好地表现其逐步增长的过程,并且能够明显地表现不同算法在训练数据集规模增加的过程中,算法精度提高的速度差异。

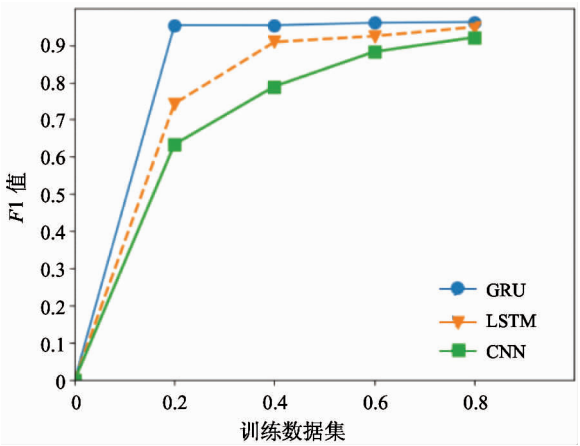

图3　3个模型*F*1值的变化情况

3.3.2　用户粉丝量实验结果分析

使用5折交叉验证法,计算预测用户粉丝量实验的RMSE与RAE,结果如表6所示。

表6　用户粉丝量预测算法准确度比较

模型	RMSE	RAE
无隐层	0.97	0.22
1层隐层	0.91	0.17
2层隐层	0.90	0.16
3层隐层	0.87	0.12

从表6的实验结果可以看出,在3种模型中,随着神经网络隐含层数量的增加,RMSE和RAE逐步减少,这表明隐含层越多预测效果越好。

4　结　论

指标可信度差异严重影响舆情指标体系的构建和应用,本文针对该问题,综合多变量数据拟合、主成分分析、多输出神经网络等技术,以及基于数据类型的指标评价方法,设计舆情指标拟合优化模型,并采用新浪微博用户数据进行性别判断实验与用户粉丝量实验,证明本方法的实用性。准备采用更多的实际数据进行测试,从而不断完善和优化本方法。在后继工作中,还需要研究指标所在层次结构对拟合效果的影响。此外,在本文方法基础上,如何通过原始数据的合理加工,构建粒度科学、实用的舆情指标体系,也是未来的研究方向。

参考文献

[1] 王光. 公安工作评价指标体系研究[C]. 见:第三届中国公共服务评价国际研讨会论文集, 中国,成都, 2011. 211-241

[2] 曾润喜, 徐晓林. 网络舆情突发事件预警系统、指标与机制[J]. 情报杂志, 2009, 28(11): 51-54

[3] 林琛. 基于网络舆论形成过程的舆情指标体系构建研究[J]. 情报科学, 2015, 33(1):146-149,161

[4] 谢海光, 陈中润. 互联网内容及舆情深度分析模式[J]. 中国青年政治学院学报, 2006, 25(3):95-100

[5] 谈国新, 方一. 突发公共事件网络舆情监测指标体系研究[J]. 华中师范大学学报, 2010, 49(3):66-70

[6] 王青, 成颖, 巢乃鹏. 网络舆情监测及预警指标体系构建研究[J]. 图书情报工作, 2011, (8):54-57

[7] 戴媛, 郝晓伟, 郭岩,等. 我国网络舆情安全评估指标体系构建研究[J]. 信息安全网络, 2010(4):12-15

[8] 张一文, 齐佳音,方滨兴,等. 非常规突发事件网络舆情热度评价指标体系构建[J]. 情报杂志, 2010, 29(11):71-76

[9] 赵建伟. 基于"用户行为"画像的实证研究——以某自营电商平台为例[J]. 科技风, 2017(22):219-221

[10] 孙新伟. 电商企业网购用户的客户分类识别研究[D]. 长春:吉林大学管理学院, 2017

[11] 姚龙飞. 基于零售信息挖掘下面向消费市场的精准推送模型设计与研究[D]. 杭州:浙江理工大学信电学院, 2017

[12] 冯娟娟, 辜丽川, 饶海笛,等. 基于客户画像和GBDT算法的客户价值预测方法[J]. 洛阳理工学院学报(自然科学版), 2018, 28(3):51-56

[13] 曾鸿, 吴苏倪. 基于微博的大数据用户画像与精准营销[J]. 现代经济信息, 2016(16):306-308

[14] 师帅, 刘沃野. 评估指标关联性因素分析方法研究[J]. 经济数学, 2010, 27(2):23-27

[15] 屈薇薇, 尚丽平, 李晓霞, 等. 基于数据拟合和主成

分分析的多组分 PAHs 神经网络定量分析[J]. 光谱学与光谱分析, 2010, 30(10):2780-2783

[16] Ding S, Su C, Yu J. An optimizing BP neural network Algorithm Based on Genetic Algorithm[J]. *Artificial Intelligence Review* (*S* 0269-2821), 2011, 36(2):153-162

[17] 刘刚. 社交媒体中微博转发的预测模型研究[D]. 北京:北京邮电大学计算机学院, 2015

[18] Chen L F, Ye Y F, Jiang Q S. A new centroid-based classifier for text categorization[C]. In: Proceeding of the IEEE 22nd International Conference on Advanced Information Networking and Applications, Okinawa, Japan, 2008. 1217-1222

[19] Zhang Y, Yang Y. Cross-validation for selecting a model selection procedure[J]. *Journal of Econometrics* (S0304-4076), 2015, 187(1):95-112

[20] GitHub Inc. 结巴中文分词[EB/OL]. https://github.com/fxsjy/jieba. GitHub Inc., 2018

Study of the fitting and optimization for public opinion analysis indicator based on the multi-output neural network

Chen Juan[*], Wang Gongming[**], Xu Yilong[***], Wang Haiwei[****]

([*] School of Journalism and Communication, Peking University, Beijing 100190)

([**] Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101)

([***] Smart City College, Beijing Union University, Beijing 100101)

([****] Military Commission Logistics Support Information Center, Beijing 100842)

Abstract

The establishment of a complete Internet public opinion indicator system based on Internet media data is the basis for public opinion prediction and evaluation as well as cyberspace governance. However, due to data conflicts, incomplete data, calculation errors, labeling errors and other problems, the credibility of some indicators is seriously reduced. In this paper, public opinion indicators are divided into two categories according to the level of credibility. And the credibility of low reliability indicator can be improved according to the high reliability one by integrating multi-variable data fitting, principal component analysis, multi-output neural network and other technologies, as well as the indicator evaluation method based on data type. In addition, the gender judgment experiment and user followers experiment are carried out with the sinaweibo user data. The experiment results show that the gender accuracy rate is as high as 96.7%, and relative absolute error (RAE) of user followers is 16%, which indicates that the proposed method can build a highly credible public opinion indicator system and lay a foundation for the establishment and quantitative research of the public opinion indicator system.

Key words: public opinion indicator system, credibility, indicator fitting, principal component analysis (PCA), multi-output neural network