

基于 SVM 的视频用户需求预测算法

江翠丽, 曹腾飞*, 李长哲, 王晓英

(青海大学 计算机技术与应用系, 青海 西宁 810016)

摘要:移动互联网的不断成熟和智能终端设备的普及推动了视频行业的发展, 视频用户喜好的不同使得用户对视频内容的需求也不相同, 精准的用户需求预测有利于从大量视频中寻找用户所需的视频内容, 节约用户寻找其喜爱视频内容的时间, 提升用户体验。为实现用户需求预测, 该文基于支持向量机(support vector machine, SVM)理论构建不同核函数下的用户需求预测模型, 并利用迭代特征消除法对小波去噪后的数据进行降维处理, 降低预测模型的复杂度。通过对不同核函数下 SVM 用户需求预测模型的对比分析, 得到参数优化后的 SVM 用户需求预测模型具有最佳的预测效果, 验证了参数优化对预测模型的重要性。优化后的 SVM 用户需求预测模型的预测结果均方差(mean square error, MSE)为 0.012 5, 平均绝对百分比误差(mean absolute percentage error, MAPE)为 0.075 5, 准确率(Accuracy)为 90.23%。

关键词:支持向量机;核函数;用户需求预测;小波去噪;迭代消除法

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2021)05-0038-05

doi:10.3969/j.issn.1673-629X.2021.05.007

Video User Demand Prediction Algorithm Based on SVM

JIANG Cui-li, CAO Teng-fei*, LI Chang-zhe, WANG Xiao-ying

(Department of Computer Technology and Applications, Qinghai University, Xining 810016, China)

Abstract: The continuous maturity of the mobile Internet and the popularization of smart terminal devices have promoted the development of the video industry. Different video users have different preferences and different demands for video content. Accurate user demand prediction helps to find the video content that users need from a large number of videos, thereby saving the time to find their favorite video content and improving the user experience. In order to realize the user demand prediction, based on the theory of support vector machine (SVM), an user demand prediction model under different kernel functions is constructed and the iterative feature elimination method is used to reduce the dimensionality of the data after wavelet denoising to reduce the complexity of the prediction model. By comparing and analyzing SVM user demand prediction models under different kernel functions, it is obtained that the parameter optimization SVM user demand prediction model has the best prediction effect, thereby the essence of parameter optimization on the prediction model is verified. The optimized SVM user demand prediction model has a mean square error (MSE) of 0.012 5, an average absolute percentage error (MAPE) of 0.075 5 and an accuracy rate of 90.23%.

Key words: support vector machine; kernel function; user demand prediction; wavelet denoising; iterative elimination method

0 引言

随着移动互联网技术的飞速发展, 视频点播、视频聊天等流媒体网络业务已成为用户消费的主流^[1]。据中国互联网络信息中心(CNNIC)第44次《中国互联网络发展状况统计报告》统计, 截至2019年6月, 国内网络视频用户规模达8.47亿, 较2018年底增长3391万, 占网民整体的88.8%^[2]。为实现高带宽、大容量、低时延, 5G移动通信网络应运而生^[3]。5G网络的超高传输速率虽然解决了网络带宽与时延的问题, 但仍

无法满足不同用户对不同视频业务的体验需求^[4]。因此, 从用户需求感知出发, 提升用户的服务体验是网络视频行业发展的必然趋势。精准的用户需求预测可以减少用户寻找所需视频内容的时间, 在提升用户体验的同时有助于运营商精准地掌握用户喜好信息, 促进网络平台的发展。

文献[5-6]将网络负荷作为评价指标构建了基于用户感知的网络负荷模型, 通过对不同小区的用户使用流量情况分析, 从用户感知速率角度实现了流量增

收稿日期:2020-06-14

修回日期:2020-10-16

基金项目:国家自然科学基金资助项目(61762074);青海省自然科学基金资助项目(2020-ZJ-943Q)

作者简介:江翠丽(1995-),女,硕士研究生,研究方向为数据挖掘;通信作者:曹腾飞(1987-),男,博士,讲师,CCF会员(34077M),研究方向为网络服务推荐与管理。

长与网络负载能力的平衡。文献[7]为应对不同用户的不同业务需求,提出了一种面向排名的预测方法,通过考虑用户对服务质量的态度和期望,来提高服务等级预测的准确性,发现满意度更高的云服务候选者,以减少负面顾客在排名相似度计算中的影响。文献[8]提出了一种基于卷积神经网络算法构建用户感知评价模型,通过对产品的使用数据进行研究分析,建立了用户感知评估与产品性能之间的映射关系,从用户感知产品性能的角度出发,预测出影响用户感知的产品性能参数。文献[9]从网络业务角度出发,充分考虑不同业务类型对信道资源的占用情况,根据流量与用户体验的变化率来构造网络效益函数,实现无线资源的利用效率。文献[10]基于深度神经网络 DNN 模型对 IPTV 视频用户的点播行为进行分析,如用户的在线时长、观看时长等,目的在于帮助 IPTV 服务提供商进行合理的资源分配,提升对用户的服务质量。

现有的视频用户需求感知方法仅依赖于用户对视频播放质量和网络状态需求的感知,忽略了用户自身观看喜好和观看行为对需求感知的重要性,而该文从用户观看视频行为出发,结合用户观看视频喜好,提出了基于 SVM 的视频用户需求预测模型,目的在于精准预测用户对视频内容的需求,帮助运营商提高视频推荐成功率,在促进网络视频业务发展的同时减少用户搜索视频的时间,提升用户体验。

1 相关理论

用户观看视频的行为受到多种因素的影响,如地区、年龄、终端设备能耗及视频的流行度等因素,不同的地区、不同的人群观看视频的行为具有显著差异^[11]。面对海量的网络视频资源,不同的用户对视频内容的需求也不同,挖掘用户观看视频行为数据中各用户与其对视频需求的关系并建立有效的视频用户需求预测模型,可以实现用户对视频内容需求的预测,提升用户体验。

由已知的用户观看视频行为数据来预测用户后续的需求非常符合支持向量机(support vector machine)的应用场景。支持向量机是由 Cortes 和 Vapnik 于 1995 年首先提出的建立在统计学习理论的 VC 维理论。该理论可以根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳解决策略,以期获得最小化结构风险^[12]。SVM 预测模型的主要思想是实现样本数据的线性回归,具体地,将线性不可分的样本数据 x 通过非线性映射 φ ,映射到更高维的特征空间 F ,在高维空间 F 实现样本数据的线性回归。由于 SVM 可以实现对特定训练样本的学习精度和无错误地识别任意样本能力的折中,从而被广泛应用于函数拟合等其

他机器学习的问题中^[13]。

为了检测由巨大的网络流量数据引起的网络入侵,文献[14]针对非线性海量数据提出了一种基于深度置信网络和支持向量机(DBN-SVM)的分类算法,提高了分类的准确性,并为实时的网络入侵检测提供了基础。文献[15]将 SVM 预测模型应用于经济预测领域,研究结果表明,支持向量机可以有效地减少噪声数据的影响,并实现非线性区域经济变量之间的映射,提高了模型的预测精度。

实现用户需求预测的目的在于:当用户浏览视频网站时,运营商可以基于用户需求预测结果为目标用户提供其可能感兴趣的视频内容,从而在提升用户观看体验的同时提高视频资源的利用率,降低网络运营成本。该文基于 SVM 预测模型实现对用户观看视频需求的预测,主要包括以下 5 个步骤:先获取用户信息及用户观看视频信息;其次对获取的用户和视频数据进行探索及处理;然后利用训练数据训练模型并生成预测模型;再利用测试数据进行模型预测;最后将预测结果与实际情况进行对比分析,实现对模型性能的评估。

2 SVM 算法

2.1 问题描述

构建基于 SVM 的视频用户需求预测模型的关键在于寻求一条泛化性比较好的决策边界,使得支持向量距离决策边界尽可能的远。寻求最优决策边界的过程即为最大化支持向量到决策边界的距离 d 的过程。

定义样本的个数为 $M = \{1, 2, \dots, m\}$,假设 x 表示原来的样本点,用 $\varphi(x)$ 表示 x 映射到高维特征空间 F 后得到的新向量,那么支撑向量到超平面 $w^T \varphi(x) + b = 0$ 的几何间隔表示为:

$$d = \frac{|w^T \varphi(x) + b|}{\|w\|} \quad (1)$$

则优化目标表示为:

$$\max d = \frac{|w^T \varphi(x) + b|}{\|w\|} \quad (2)$$

注意到几何间隔 d 与 $\|w\|$ 是成反比的,且 $|w^T \varphi(x) + b|$ 为固定值,因此最大化几何间隔 d 的问题可以转化为最小化 $\|w\|$ 的问题。为了方便后边求导计算,将目标优化问题(2)进一步转化为(3):

$$\min \frac{1}{2} \|w\|^2 \quad (3)$$

为提高 SVM 视频用户需求预测模型的泛化能力,引入松弛变量 ξ_i ,则目标优化问题将进一步转化为:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$$

$$\begin{aligned} \text{s. t. } & 1 - y_i(\mathbf{w}^T \boldsymbol{\varphi}(x_i) + b) - \xi_i \leq 0 \\ & \xi_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (4)$$

其中, C 为惩罚因子且满足 $C > 0$ 。分析可知: 为使得样本数据线性可分, 需满足 C 为无穷大时, ξ_i 必然无穷小; 当且仅当 C 为有限值的时候, 才会允许部分样本不遵循约束条件。

2.2 优化目标求解

通过添加拉格朗日乘子, 构造拉格朗日函数, 将问题(4)转化为无约束的优化问题, 如下:

$$\begin{aligned} \min_{w, b, \xi} \max_{\lambda, \mu} L(w, b, \xi, \lambda, \mu) &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i + \\ & \sum_{i=1}^n \lambda_i [1 - \xi_i - y_i(\mathbf{w}^T \boldsymbol{\varphi}(x_i) + b)] - \sum_{i=1}^n \mu_i \xi_i \\ \text{s. t. } & \lambda_i \geq 0, \mu_i \geq 0 \end{aligned} \quad (5)$$

其中, λ_i 和 μ_i 为拉格朗日乘子, 利用强对偶性转化, 将式(5)转化为:

$$\begin{aligned} \max_{\lambda, \mu} \min_{w, b, \xi} L(w, b, \xi, \lambda, \mu) &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i + \\ & \sum_{i=1}^n \lambda_i [1 - \xi_i - y_i(\mathbf{w}^T \boldsymbol{\varphi}(x_i) + b)] - \sum_{i=1}^n \mu_i \xi_i \\ \text{s. t. } & \lambda_i \geq 0, \mu_i \geq 0 \end{aligned} \quad (6)$$

分别对 \mathbf{w} , b 和 ξ_i 求偏导数, 并令偏导数为 0, 得出如下关系:

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = \sum_{i=1}^m \lambda_i \boldsymbol{\varphi}(x_i) y_i - \mathbf{w} = 0 \\ \frac{\partial L}{\partial b} = \sum_{i=1}^m \lambda_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = \lambda_i + \mu_i - C = 0 \end{cases} \quad (7)$$

将式(7)中的求导结果带入式(6), 得到新的目标优化函数:

$$\begin{aligned} \max_{\lambda} & \left[\sum_{j=1}^n \lambda_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j (\boldsymbol{\varphi}(x_i) \cdot \boldsymbol{\varphi}(x_j)) \right] \\ \text{s. t. } & \sum_{j=1}^n \lambda_j y_j = 0, \lambda_i \geq 0, C - \lambda_i - \mu_i = 0 \end{aligned} \quad (8)$$

求解式(8), 可以得到:

$$\mathbf{w} = \sum_{i=1}^m \lambda_i \boldsymbol{\varphi}(x_i) y_i \quad (9)$$

将支持向量 (x_s, y_s) 带入: $y_s(\mathbf{w}^T \boldsymbol{\varphi}(x_s) + b) = 1$, 求出 b 为:

$$b = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} (y_s - \mathbf{w}^T \boldsymbol{\varphi}(x_s)) \quad (10)$$

其中 $|\mathcal{S}|$ 为支持向量的个数。求出 \mathbf{w} 和 b 后, 便可以求得超平面 $\mathbf{w}^T \boldsymbol{\varphi}(x) + b = 0$ 。

3 数据处理

3.1 数据集

文中用到的数据是利用 Python 软件爬取的哔哩

哔哩网站上的用户观看视频行为信息。首先爬取哔哩哔哩网站上热门番剧的 ID, 然后根据番剧 ID 爬取番剧信息及追番用户的相关信息。最终用于实验仿真的数据集是包含 10 000 行 12 列的用户-视频信息数据。

3.2 消噪

为避免噪声数据对数据体量、复杂度以及处理结果准确率和时效性的影响, 利用小波方法对样本数据进行消噪处理, 消除数据集中无效或异常的数据, 提高数据的收敛速度和模型预测的精度^[16-18]。

由图 1 可以看出, 经小波去噪后的数据与原数据相比, 数据的大体趋势并未发生变化。小波消噪可以在保证数据连续性和科学性的前提下, 避免噪声数据对预测模型的扰动。

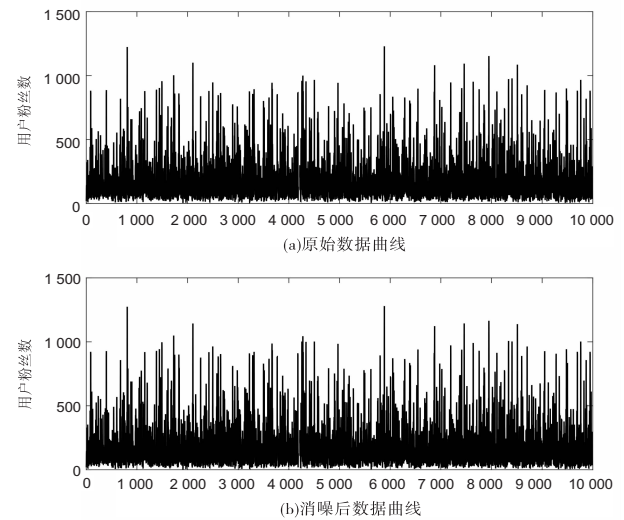


图 1 小波消噪前后对比

3.3 降维

为避免多个属性间强相关性对数据集分析产生干扰, 提升预测模型的可靠性, 需从番剧标签、追番人数、番剧评分、用户性别、用户等级、用户粉丝数、用户关注人数、用户评论点赞数这 8 个属性标签中选取对预测用户评分贡献较大的特征属性, 降低预测模型的复杂度。在对数据集进行降维处理的过程中, 需要保证筛选出的属性要尽可能多地表征样本数据的大部分信息, 否则将不会降低数据分析的难度和复杂度, 甚至会导致分析失败。

将番剧标签、追番人数、番剧评分、用户性别、用户等级、用户粉丝数、用户关注人数、用户评论点赞数这 8 个属性标签分别编号为 1-8, 通过递归特征消除法对样本数据中各属性的贡献值进行筛选^[19]。

图 2 为各属性对预测结果的贡献值, 设定阈值 0.85 可以将原来的 8 维数据集降为 5 维数据集。选出的 5 个主要特征属性分别为: 番剧标签、追番人数、番剧评分、用户性别、用户等级。将筛选出的 5 个属性用于模型训练。

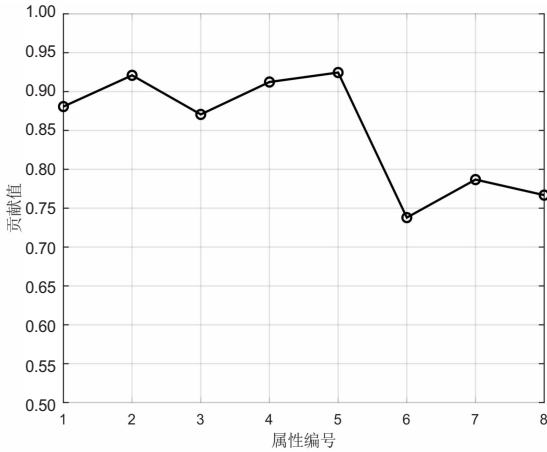


图2 各属性贡献值

3.4 归一化

为避免因属性量化级不同带来的预测偏差,利用 min-max 标准化公式对样本数据进行归一化处理。

$$x_{inorm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (11)$$

其中, x_i 为属性中第 i 个样本的数据; x_{max} 和 x_{min} 为属性的最大值和最小值。

图3 为数据集中原始的番剧评分数据与归一化处理后的番剧评分数据对比图。归一化处理避免了因样本数据各属性间量化级不同造成的误差,可以提升模型的收敛速度和精度。

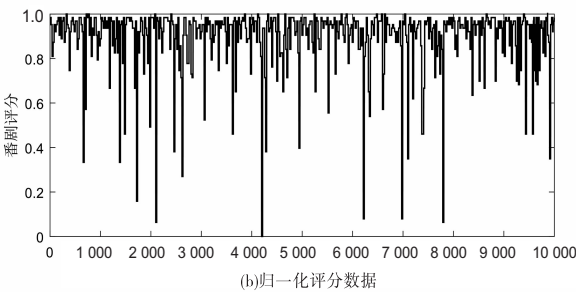
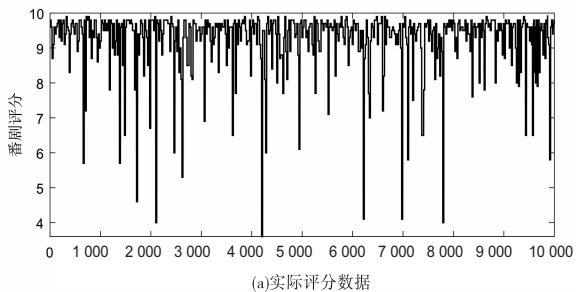


图3 归一化处理前后对比

4 实验与分析

4.1 基于不同核函数的 SVM 视频用户需预测模型

首先将数据集分为训练集和预测集两类,将数据集的 70% 作为训练集,30% 作为测试集。在模型训练的过程中,分别基于表 1 中三个不同的核函数构造 SVM 视频用户需求预测模型,预测结果如图 4 所示。

表1 不同核函数的表达式

名称	表达式	参数
线性核	$K(x_i, x_j) = x_i^T x_j$	/
多项核	$K(x_i, x_j) = (x_i^T x_j)^d$	$d \geq 0$ 为多项式的次数
高斯核	$K(x_i, x_j) = \exp(-\delta \ x_i - x_j\ ^2)$	$\delta > 0$ 为高斯核的带宽

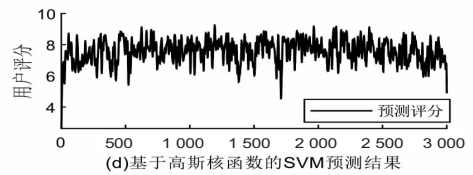
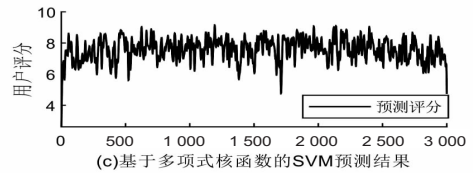
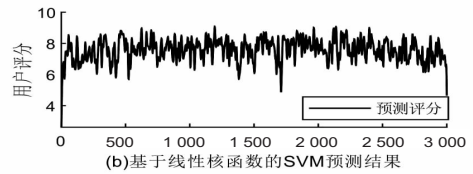
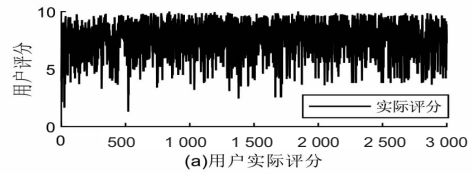


图4 基于不同核函数的 SVM 预测结果对比

选取均方误差 (MSE)、平均绝对百分比误差 (MAPE) 和准确率 (Accuracy) 三个评价指标分别从预测误差和预测精准度两个方面对不同核函数下的 SVM 视频用户需求预测结果进行对比,结果如表 2 所示。

表2 基于不同核函数的 SVM 用户需求预测效果对比

核函数	MSE	MAPE	Accuracy/%
线性核	0.020 8	0.123 2	71.23
多项核	0.020 4	0.121 4	78.33
高斯核	0.019 9	0.118 2	81.48

根据统计学原理可知,MSE、MAPE 两者的数值越小,表示预测值与真实值的误差越小,即预测结果越接近于真实值,而准确率越高表明预测模型的预测效果越好。对比表 2 中不同核函数下的 SVM 预测结果可知,基于高斯核函数的 SVM 视频用户需求预测模型预测效果较好。

4.2 参数调优

为降低视频用户需求预测模型的预测误差,提高预测的准确率,现针对高斯核函数下的 SVM 视频用户需求预测模型进行参数调优。随机选取 4 组参数对 (C, δ) 进行对比实验,其中 C 为惩罚系数, δ 为高斯核

函数系数,预测对比结果如图 5 所示。

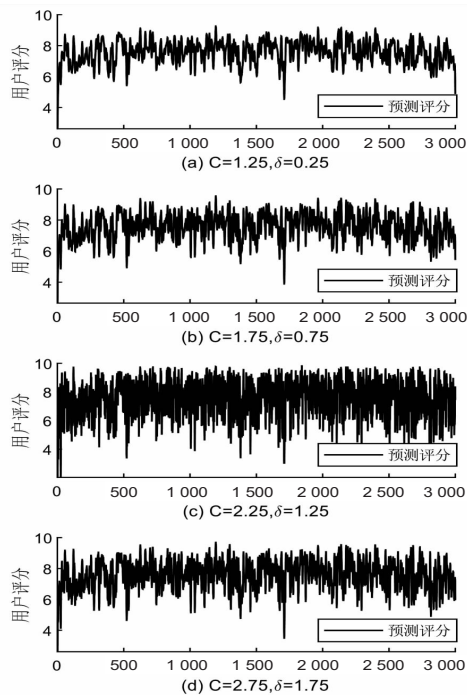


图 5 基于不同参数对的 SVM 预测结果对比

通过对比表 3 中不同参数对 SVM 视频用户预测模型的 MSE、MAPE 和 Accuracy,可以得到 $C = 2.25$, $\delta = 1.25$ 时的 SVM 视频用户需求预测模型最佳。

表 3 基于参数 C 和 δ 的 SVM 预测效果对比

参数	MSE	MAPE	Accuracy/%
$C = 1.25, \delta = 0.25$	0.019 9	0.118 2	85.89
$C = 1.75, \delta = 0.75$	0.018 0	0.106 6	87.42
$C = 2.25, \delta = 1.25$	0.012 5	0.075 5	90.23
$C = 2.75, \delta = 1.75$	0.016 2	0.096 1	89.87

5 结束语

文中将用户观看视频行为的数据作为训练集,以用户数据及其对视频需求的关系为基准,提出基于不同核函数下的 SVM 视频用户需求预测模型,并对高斯核函数下的 SVM 视频用户需求预测模型进行参数调优,通过对预测结果的误差和准确度进行对比分析,验证了参数优化的有效性。优化后的 SVM 视频用户需求预测模型的预测结果的 MSE 为 0.012 5, MAPE 为 0.075 5, 准确度为 90.32%。该研究成果有助于运营商精准掌握用户需求,提升平台为用户推荐视频的成功率,减少用户在海量网络视频资源中搜索视频内容的时间。未来还需进一步考虑如何结合神经网络算法对参数进行智能调优,进一步降低模型预测的误差,以提高预测的准确度。

参考文献:

[1] 廖秉宜,姜佳妮.中国网络视频产业组织优化与规制政策

研究[J].新闻与传播评论,2019,72(3):64-74.

- [2] 郭加佳.“二更”视频的内容生产与传播创新[J].传媒,2019(11):52-53.
- [3] JIANG C, CAO T, LI C, et al. Multimedia resources purchase and distribution optimization for 5G edge cloud networks[J]. IEEE Access, 2020, 8: 59340-59350.
- [4] 孙琦,陈桂芬.5G中基于穷举搜索的D2D资源分配算法[J].计算机应用研究,2019,36(11):3456-3459.
- [5] 刘毅,刘立洋,吴德胜.基于用户感知的网络负荷研究[J].山东通信技术,2019,39(1):25-28.
- [6] 于德成,薛振,丁章,等.LTE网络用户感知速率提升策略研究[J].电信工程技术与标准化,2019,32(8):69-73.
- [7] DING Shuai, WANG Zeyuan, WU Desheng, et al. Utilizing customer satisfaction in ranking prediction for personalized cloud service selection[J]. Decision Support Systems, 2017, 93: 1-10.
- [8] 颜波,张磊,褚学宁.基于卷积神经网络的用户感知评估建模[J].上海交通大学学报,2019,53(7):844-851.
- [9] 张之栋,李威,陈洁,等.基于用户感知和网络效益的LTE无线扩容方法[J].电信工程技术与标准化,2019,32(10):46-50.
- [10] 刘超,贾毓臻,王攀.基于深度学习的IPTV视频用户行为分析方法研究[J].计算机应用与软件,2019,36(6):167-170.
- [11] 寇兰,刘彪,黄宏程.用户需求感知的D2D视频分发机制[J].计算机应用研究,2019,36(6):1849-1853.
- [12] 张益铭,徐晓钟,王智庆.支持向量机与时间序列预测综述[J].计算机应用与软件,2010,27(12):127-129.
- [13] 杨永琪,徐欣蕾,陈曦.基于SVM技术实现手写数字分类识别的研究[J].电脑知识与技术,2020,16(6):195-196.
- [14] ZHANG Hao, LI Yongdan, LV Zhihan, et al. A real-time and ubiquitous network attack detection based on deep belief network and support vector machine[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(3): 790-799.
- [15] ZHANG Kun. Forecasting regional economic growth using support vector machine model[J]. Ecological Economy, 2019, 15(3): 186-192.
- [16] XIAO M, WEN K, ZHANG C, et al. Research on fault feature extraction method of rolling bearing based on NMD and Wavelet threshold denoising[J]. Shock and Vibration, 2018, 2018 (PT. 7): 9495265. 1-9495265. 11.
- [17] 李佩钰.一种基于小波和神经网络的短时交通流量预测[J].计算机技术与发展,2020,30(1):135-139.
- [18] 董胡,徐雨明,马振中,等.基于小波包与自适应维纳滤波的语音增强算法[J].计算机技术与发展,2020,30(1):50-53.
- [19] 申滨,赵书锋,金纯.基于迭代并行干扰消除的低复杂度大规模MIMO信号检测算法[J].电子与信息学报,2018,40(12):2970-2978.