

基于聚类分析算法和优化支持向量机的无线网络流量预测



曹素娥 杨泽民

山西大同大学计算机与网络工程学院 山西 大同 037009

摘要 为了解决当前无线网络流量预测过程存在的一些问题,以提高无线网络流量的预测精度为目标,提出基于聚类分析算法和优化支持向量机的无线网络流量预测模型。首先,采集无线网络流量数据集,并采用聚类分析算法构建训练样本集合;然后,采用支持向量机对无线网络流量训练样本进行学习,并引入布谷鸟搜索算法对支持向量参数进行优化,从而建立无线网络流量预测模型;最后,通过具体无线网络流量预测实例分析模型的有效性。结果表明,所提模型的无线网络流量预测精度高,提升了无线网络流量建模效率,而且其无线网络流量预测效果要优于当前经典无线网络流量预测模型,具有比较显著的优越性。

关键词 无线网络;流量预测;聚类分析算法;支持向量机;布谷鸟搜索算法

中图分类号 TP391

Prediction of Wireless Network Traffic Based on Clustering Analysis and Optimized Support Vector Machine

CAO Su-e and YANG Ze-min

School of Computer and Network Engineering, Shanxi Datong University, Datong, Shanxi 037009, China

Abstract In order to solve the problems existing in the current wireless network traffic prediction process and improve the accuracy of wireless network traffic prediction, a wireless network traffic prediction model based on clustering analysis algorithm and optimized support vector machine is proposed. Firstly, data sets of wireless network traffic are collected and clustering analysis algorithm is used to construct the training sample set. And then, support vector machine is used to learn the training samples of wireless network traffic, and cuckoo search algorithms is introduced to optimize the parameters of support vector. Thus, the prediction model of wireless network traffic is established. Finally, the effectiveness of the model is analyzed through a specific example of wireless network traffic prediction. The results show that the proposed model has high prediction accuracy, improves the efficiency of wireless network traffic modeling, and the prediction effect of wireless network traffic is better than the current classical wireless network traffic prediction models, which has more significant advantages.

Keywords Wireless network, Traffic prediction, Clustering analysis algorithm, Support vector machine, Cuckoo search algorithms

1 引言

随着移动技术、传感器技术、无线通信技术和微电子技术等不断发展和融合,各种类型的无线通信网络系统随之出现,它们给人们的工作和生活带来便利^[1-3]。随着无线网络用户不断增多,每天无线网络上的数据量不断增加,尤其是节假日、周末以及晚上等上网高峰期,无线网络堵塞的概率相当高,严重影响人们正常上网^[4-6]。无线网络流量预测可以帮助管理人员和网络公司了解无线网络流量的变化特点,及时跟踪网民上网的行为特征,提前制定相应的无线网络管制方案,以防止出现无线网络堵塞的现象。因此无线网络流量预测成为无线网络管理研究中的一个重要方向,对其进行研究具有重要的理论价值和实际应用价值^[7]。

由于无线网络是一个新事物,其流量的预测基本沿用有线网络的流量建模方法。传统无线网络流量预测模型主要为时间序列分析法,该方法首先将无线网络流量历史样本按时间先后顺序进行排列,选择一部分无线网络流量历史样本作为训练样本,对无线网络流量预测模型的参数进行估计,并且对另一部分无线网络流量历史样本进行预测,对预测效果进行统计和分析,但该方法主要针对小规模无线网络,不能满足现代无线网络向大规模方向发展的要求^[8-10]。现代无线网络流量预测模型采用非线性建模技术,如BP神经网络的无线网络流量预测模型、RBF神经网络的无线网络流量预测模型、支持向量机的无线网络流量预测模型等,它们具有较强的自适应学习能力,可以对无线网络流量变化特点进行拟合,无线网络流量预测效果要优于传统模型^[10-12]。训练样本集合

到稿日期:2019-08-16 返修日期:2019-10-17 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(11871314);国家自然科学基金青年基金(11605107)

This work was supported by the National Natural Science Foundation of China (11871314) and National Natural Science Youth Foundation of China (11605107).

通信作者:曹素娥(zl201851402@yeah.net)

的构建对无线网络流量预测结果具有决定性的影响,当前主要根据时间前后,或采用随机方式选择训练样本集合。无法建立性能最优的无线网络流量预测模型,就无法获得高精度的无线网络流量预测结果,同时会延长建模时间,影响无线网络流量预测建模的效率^[13-15]。

针对当前无线网络流量预测误差大、预测结果可信度低等问题,为了获得更优的无线网络流量预测结果,本文设计了聚类分析和优化支持向量机的无线网络流量预测模型,采用聚类分析算法构建最优的训练样本,然后引入优化支持向量机建立无线网络流量预测模型。通过与其他无线网络流量预测模型进行对比测试,验证了本文无线网络流量预测建模的可行性,并且其预测结果要明显优于其他模型,表明该模型是一种精度高、效率高的无线网络流量预测方法。

2 聚类分析和优化支持向量机的无线网络流量预测模型

2.1 聚类分析算法

当前有很多类型的聚类分析算法,如 k-均值聚类分析算法、模糊均值聚类算法,其中模糊均值聚类算法的通用性强,性能更优,因此本文选择模糊均值聚类算法对无线网络流量预测过程中的样本进行预处理,以选择最优的训练样本集合^[16]。

设无线网络流量的数据集为 $X = \{x_1, x_2, \dots, x_n\}$, n 表示样本的数量,则该无线网络流量的数据集可以划分为 c 类,它们的聚类中心集合为 $V = \{v_1, v_2, \dots, v_c\}$,每一个无线网络流量均可以划分到相应的类别中,每一类都有一个隶属度值 u_{ij} ,它们构成一个隶属度矩阵 $U = (u_{ij})_{c \times n}$,模糊均值聚类算法的优化目标函数为:

$$J_m(U, v_1, v_2, \dots, v_c) = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m d_{ij}^2 \quad (1)$$

其中, m 表示模糊系数, d_{ij} 表示样本 x_j 与聚类中心 v_i 之间的距离。

由于 $\sum_{i=1}^c u_{ij} = 1$,采用拉格朗日乘算法对最优聚类中心进行求解,得到:

$$F = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^m d_{ij}^2 + \lambda(1 - \sum_{i=1}^c u_{ij}) \quad (2)$$

对式(2)进行求偏导,得到 u_{ij} 的计算式为:

$$u_{ij} = \frac{1}{\sum_{k=1}^c (d_{ij}/d_{kj})^{\frac{2}{m-1}}} \quad (3)$$

v_i 的计算式为:

$$v_i = \sum_{j=1}^n u_{ij}^m x_j \times \frac{1}{\sum_{j=1}^n u_{ij}^m} \quad (4)$$

2.2 优化支持向量机

2.2.1 支持向量机

设训练样本 $T = \{(x_1, y_1), \dots, (x_n, y_n)\}$, n 为训练样本数,支持向量机回归可以描述为一个带约束条件的二次规划问题求解,即有:

$$\begin{aligned} \min & \frac{1}{2} (\omega \cdot \omega^T) + C \sum_{i=1}^n \varepsilon_i \\ \text{s. t. } & y_i [(\omega \cdot x_i) + b] \geq 1 - \varepsilon_i \\ & \varepsilon_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (5)$$

其中, C 为惩罚因子。

引入 Lagrange 乘子 $\alpha_i (i = 1, \dots, n)$ 建立式(5)的对偶形式,具体为:

$$\begin{aligned} \max & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n y_i y_j \alpha_i \alpha_j \\ \text{s. t. } & 0 \leq \alpha_i \leq C, i = 1, 2, \dots, n \\ & \sum_{i=1}^n y_i \alpha_i = 0 \end{aligned} \quad (6)$$

对于非线性回归问题,引入核函数 $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ 进行映射,则式(4)可以为:

$$\max \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n y_i y_j \alpha_i \alpha_j k(x_i, x_j) \quad (7)$$

支持向量机回归的决策函数为:

$$f(x) = \sum_{i=1}^n y_i \alpha_i k(x_i, x_j) + b \quad (8)$$

其中, b 是阈值。核函数采用 RBF 核函数,具体如下:

$$k(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2} \quad (9)$$

其中, σ 表示核宽度参数。

2.2.2 布谷鸟搜索算法

布谷鸟搜索算法源于布谷鸟种群寄生繁衍机制,其利用果蝇的莱维飞行模式,具有较强的智能搜索优势。该算法设置了3个规则,具体如下:

(1) 每只布谷鸟每次只下一个蛋,该蛋随机置于一个鸟巢,与问题的一个解相对应;

(2) 部分鸟巢放着优质蛋,即问题的较优解,这些鸟巢会直接传到下一代;

(3) 布谷鸟被寄主发现的概率为 Pa ,如果某个鸟巢被发现,则该鸟巢被抛弃,为了保证鸟巢数量不变,需要产生相同数量的新鸟巢。

2.2.3 布谷鸟搜索算法优化支持向量机参数

支持向量机主要待优化的参数为 C 和 σ ,它们过大或过小,均会产生过拟合或欠学习的问题,为此本文选择布谷鸟搜索算法优化支持向量机参数,具体步骤如下。

Step1 设置训练样本集合和测试样本集合,设置 C 的取值范围。

Step2 始化 m 个两维的鸟巢 $D = \{x_i\} = \{[C, \sigma]\}, i = 1, 2, \dots, m$,计算每一组鸟巢的适应度函数值 F_i 。

Step3 设置迭代次数 $t = 0$ 。

Step4 根据适应度函数值对鸟巢优劣进行排序。

Step5 设置 Pa 的值,确定需要抛弃的鸟巢,根据莱维飞行可得一些新的鸟巢,用这些新鸟巢代替抛弃的鸟巢,并计算新鸟巢的适应度值。

Step6 迭代次数增加。

Step7 如果迭代次数等于最大迭代次数,则根据最优鸟巢得到最优参数 C 和 σ ,否则返回 Step4 继续迭代。

2.3 聚类分析算法和优化支持向量机的无线网络流量预测步骤

(1) 对于某个无线网络通信系统,采集其一段时间的流量数据,并补全丢失的数据。

(2) 采用模糊聚类分析算法对无线网络流量数据进行聚类操作,选择与每一个无线网络预测样本点相关的训练样本集合。

(3) 设置支持向量的参数取值范围,并初始化核宽度和惩罚参数。

(4)采用支持向量机对无线网络流量训练样本进行学习。

(5)采用布谷鸟搜索算法对核宽度和惩罚参数进行优化。

(6)统计无线网络流量训练误差,如果误差小于无线网络管理的实际要求范围,则得到最优核宽度和惩罚参数的值,否则返回步骤(4)继续进行学习。

(7)支持向量机训练终止,根据最优核宽度和惩罚参数建立无线网络流量的预测模型。

基于聚类分析和优化支持向量机的无线网络流量预测流程如图1所示。

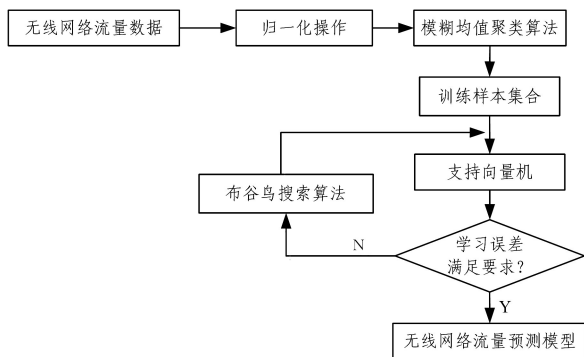


图1 聚类分析和优化 SVM 的无线网络流量预测流程

Fig. 1 Flow of wireless network traffic prediction based on clustering analysis and SVM

3 无线网络流量预测的实例分析

3.1 实验数据

为了分析基于聚类分析和 SVM 的无线网络流量预测模型的效果,本节进行实验验证。实验数据来源于网络流量文库^[17],共收集 1 200 个无线网络流量数据作为研究对象,它们的变化曲线如图 2 所示。

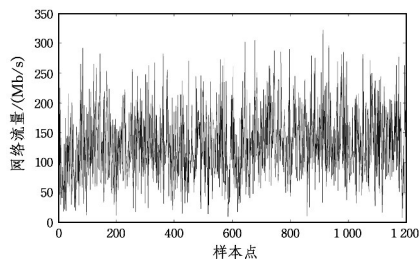


图2 1200个无线网络流量数据

Fig. 2 1200 wireless network traffic data

从图2可以看出,无线网络流量数据之间的值变化较大,会对支持向量机的训练产生不利影响,为此对其进行如下处理^[18]:

$$x'(i) = \frac{x(i) - E_x}{\sigma_x} \quad (10)$$

其中, E_x 和 σ_x 分别为均值和标准差。

本文选择基于 ARMA 的无线网络流量预测模型、BP 神经网络的无线网络流量预测模型与所提模型进行对比测试。为了体现无线网络流量预测结果的可靠性,进行 5 次仿真对比实验,每次实验的训练样本和测试样本数量如表 1 所列。

表1 无线网络流量的训练样本和测试样本数量

Table 1 Number of training and testing samples for wireless network traffic

仿真实验编号	训练样本	测试样本
1	400	800
2	600	600
3	800	400
4	1 000	200
5	1 100	100

(单位:个)

3.2 支持向量机参数的确定

每次仿真实验都通过谷鸟搜索算法优化支持向量机参数 C 和 σ ,结果如表 2 所列。可以看出,每次实验的 C 和 σ 均有一定的差别,主要是因为训练样本的规模发生了改变。

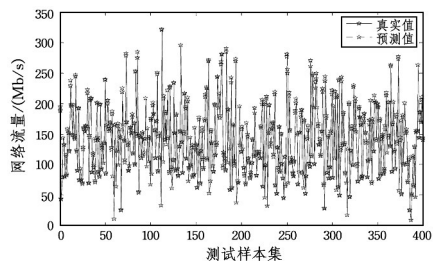
表2 不同实验的支持向量参数值

Table 2 Support vector parameter values for different experiments

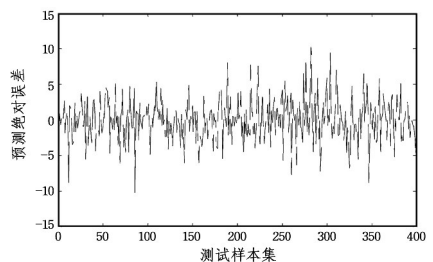
仿真实验编号	C 值	σ 值
1	27.895	5.499
2	93.781	8.806
3	82.210	2.790
4	84.592	1.860
5	37.817	6.574

3.3 无线网络流量预测的结果对比

由于篇幅有限,本文只列出训练样本为 800,测试样本为 400 时的聚类分析和 SVM 的无线网络流量预测结果,如图 3 所示。



(a) 预测值与实际值对比



(b) 预测偏差

图3 基于聚类分析和 SVM 的无线网络流量的预测结果

Fig. 3 Prediction results of wireless network traffic based on clustering analysis and SVM

统计 3 种模型的无线网络流量预测精度,如图 4 所示。可以发现,随着无线网络流量训练样本数量的不断增加,无线网络流量的预测精度不断提高,这是因为训练样本越多,无线网络流量的训练效果就越好,建立的无线网络流量预测模型越好。在训练样本数量相同的情况下,基于 ARMA 的无线

¹⁾ <http://newsfeed.ntcu.net/~news/2018/>

网络流量模型预测精度最低,因为它利用一种线性建模技术,与无线网络流量变化特点不相适应;基于 BP 神经网络的无线网络流量预测模型的精度要高于 ARMA,因为其是非线性建模技术,可以更好地跟踪无线网络流量的变化趋势;而本文模型的无线网络流量模型的预测精度最高,这是因为通过模糊聚类分析算法可以构建更优的训练样本集合,同时引入优化支持向量机进行学习,使学习性能更优,减少了无线网络流量预测误差,使得预测结果更加可靠。

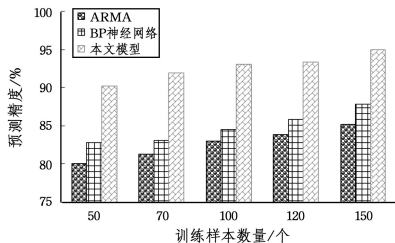


图4 无线网络流量的预测精度比较

Fig. 4 Comparisons of wireless network traffic prediction accuracy

3.4 无线网络流量的建模时间比较

统计3种模型的无线网络流量训练时间和预测时间,结果如表3所列。从表3可以看出,无线网络流量预测模型的预测时间相差不大,但是训练时间差异很大,其中基于BP神经网络的训练时间最长,因为其收敛速度慢;而本文方法的训练时间最短,因为其通过引入聚类分析算法选择最优的训练样本数据,减少了训练样本数量,降低了训练的时间复杂度,提升了无线网络流量预测的建模速度。

表3 无线网络流量预测的建模时间对比

Table 3 Modeling time comparison of wireless network traffic prediction

实验编号	ARMA		BP神经网络		本文模型	
	训练时间	测试时间	训练时间	测试时间	训练时间	测试时间
1	10.64	1.70	17.71	1.76	8.62	1.64
2	11.59	1.59	16.76	1.60	8.56	1.58
3	11.05	1.54	17.16	1.55	8.98	1.48
4	10.85	1.23	16.98	1.30	8.75	1.19
5	12.18	1.41	17.26	1.47	8.17	1.33

(单位:s)

结束语 为了解决当前无线网络流量预测过程存在的一些难题,本文采用聚类分析算法构建最优的训练样本,并采用优化支持向量机对训练样本进行学习。实验结果表明,与其他无线网络流量预测模型相比,本文模型是一种速度快、误差小的无线网络流量工具。

参考文献

[1] CINIC. The 41st China statistical report on Internet development [R]. Beijing: China Internet Network Information Center, 2018.

[2] VELAN P, CERMAK M, CELEDA P, et al. A survey of methods for encrypted traffic classification and analysis[J]. International Journal of Network Management, 2015, 25(5): 355-374.

[3] DU Z, MA L P, SUN G Z. Network Traffic Anomaly Detection Based on Wavelet Analysis [J]. Computer Science, 2019, 46(8): 178-182.

[4] GUO J, YU Y B, YANG C Y. Multi-step network traffic prediction based on full attention mechanism [J]. Signal Processing,

2019, 35(5): 758-767.

[5] CHEN X, TANG J Y. Internet of Things traffic prediction model based on Bayesian and causal ridge regression [J]. Journal of Sichuan University (Natural Science Edition), 2018, 55(5): 965-970.

[6] GUO F, CHEN L, YANG Z W. Real-time traffic prediction of large-scale IP backbone network based on MGU [J]. Journal of Shandong University (Engineering Science), 2019, 49(2): 88-95.

[7] ACETO G, CIUNZO D, MONTIERI A, et al. Mobile encrypted traffic classification using deep learning[C]// 2018 Network Traffic Measurement and Analysis Conference (TMA). IEEE, 2018: 1-8.

[8] GE S C, LIU X F, ZHOU F. Train information transmission network traffic modeling and prediction of CRH2 EMU [J]. Computer Science, 2017, 44(10): 91-95, 126.

[9] LI R, ZHAO Z, ZHENG J, et al. The learning and prediction of application-level traffic data in cellular networks [J]. IEEE Transactions on Wireless Communications, 2017, 16(6): 3899-3912.

[10] LIN Z D, LV H H. Network traffic prediction mechanism based on wavelet coefficient perception [J]. Journal of Terahertz Science and Electronic Information Technology, 2019, 17(1): 131-135.

[11] LIU Y, LI W, LI Y C. Network traffic classification using k-means clustering[C]// International Multi-symposiums on Computer & Computational Sciences, 2007.

[12] CHEN Z, LIU Z, PENG L, et al. A novel semi-supervised learning method for Internet application identification [J]. Soft Computing, 2017, 21(8): 1963-1975.

[13] LOTFOLLAHI M, SIAVOSHANI M J, ZADE R S H, et al. Deep packet: A novel approach for encrypted traffic classification using deep learning [J]. Soft Computing, 2017, 28(9): 1-14.

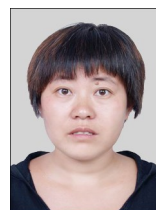
[14] LI Y Q, H Y, SUN X C. Network traffic prediction model based on deep confidence echo state network [J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2018, 38(5): 85-90.

[15] WANG K. Research on network traffic prediction based on cloud computing and extreme learning machine [J]. Journal of Shandong Agricultural University (Natural Science Edition), 2018, 49(4): 632-635.

[16] LIU K. Chaotic prediction of network traffic based on particle swarm optimization support vector machine [J]. Modern Electronics Technique, 2019, 42(2): 120-123.

[17] ZHANG C, PAUL P. Long-term mobile traffic forecasting using deep spatio-temporal neural networks[C]// Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing, Angeles: ACM, 2018: 231-240.

[18] NAREJO S, PASERO E. An Application of Internet Traffic Prediction with Deep Neural Network [J]. Multidisciplinary Approaches to Neural Computing, 2018, 69(1): 139-149.



CAO Su-e, born in 1976, master, experimenter. Her main research interests include data processing and data mining.