

基于改进极限学习机的微信热点预测

林 荫¹, 杨长春²

(1 常州大学 怀德学院, 江苏 靖江 214500;

2 常州大学 信息化建设与管理中心, 江苏 常州 213164)

摘 要: 为了提高微信热点预测的准确性, 更好描述微信热点的变化趋势, 针对当前微信热点预测模型存在的一些局限性, 设计了改进极限学习机的微信热点预测模型. 首先采集微信热点的历史数据, 构建微信热点的时间序列学习样本, 然后采用极限学习机对微信热点样本进行训练, 并对标准极限学习机进行改进, 提升其学习能力, 建立了微信热点的预测模型, 最后采用具体微信热点数据进行验证性测试. 测试结果表明, 改进极限学习机可以获得高精度的微信热点预测结果, 而且微信热点的建模效率要高于其他微信热点预测模型, 具有良好的实际应用价值.

关键词: 微信热点; 预测模型; 极限学习机; 验证性测试; 权值更新

中图分类号: TP399

文献标识码: A

文章编号: 1000—7180(2017)05—0123—05

Weixin Hot Spot Prediction by Using Modified Extreme Learning Machine

LIN Yin¹, YANG Chang-chun²

(1 School of Huaide, Changzhou University of China, Jingjiang 214500, China;

2 Information Construction and Management Center, Changzhou University of China, Changzhou 213164, China)

Abstract: In order to improve the accuracy of the weixin hot spot prediction, describe weixin hot trends, the current weixin hot forecast model has some limitations, a weixin hot prediction model based on improved extreme learning machine is proposed. Firstly, historical data are gathered for weixin hot, and time series hot study sample of weixin establisheds, and secondly, extreme learning machine was trained on weixin hot samples while standard extreme learning machine learning is modified to improve learning ability and prediction models of weixin hot is test by weixin hot data test. The test results show that modified extreme learning machine can get high accurate prediction results of weixin hot, and the modeling efficiency is higher than other models, it has good practical application value.

Key words: weixin hot spot; prediction model; extreme learning machine; verification test; weight update

1 引言

随着网络技术的不断发展, 手机成为人们主要的通信工具, 微信也成为人们交流和沟通的一种方式, 其中微信热点是人们关注的焦点. 微信热点是对某一事件的看法, 而负面的微信热点对社会的稳定具有不利影响, 因此如何对微信热点进行准确跟踪和预测是网络研究领域的重大课题之一^[1-2].

当前微信热点预测主要分为传统统计模型和现代

统计学模型, 传统统计模型主要有线性回归分析算法等^[3], 它们认为微信热点的变化趋势是一种线性规律, 而实际上微信热点受到人的思想、人们群聚, 以及外界因素的综合作用, 具有比较强烈的非线性波动特性, 这样传统模型无法跟踪微信热点的变化态势, 预测结果极不可靠^[4-5]. 现代统计学模型以非线性理论为基础, 主要以机器学习算法为主, 常用的算法有神经网络、支持向量机, 相对于线性回归算法, 它们具有更好的非线性学习和预测能力, 更加准确地描述了微信热点的动态、

时变特点,获得了更高精度的微信热点预测结果^[6-8].在实际应用中,传统神经网络、支持向量机均有自己的局限性,如支持向量机的学习迭代次数多,计算复杂度高,无法满足大规模微信热点的数据预测要求,传统神经网络在学习要不断修改权值,训练耗时相当少,微信热点预测的实时性差,影响微信热点的预测效果^[9-10].在传统神经网络的基础上,有学者提高了一种新型的神经网络—极限学习机(ELM),以广义逆矩阵理论为基础,一次完成网络学习,不需要多次迭代,计算复杂度大幅度下降,加快了学习速度,且具有支持向量机的良好泛化能力^[11-12].

为了提高微信热点预测的准确性,更好描述微信热点的变化趋势,针对当前微信热点预测模型存在局限性,设计了改进极限学习机的微信热点预测模型(modify ELM, MELM).首先构建微信热点的时间序列学习样本,然后采用 MELM 建立微信热点预测模型,并使用具体微信热点数据进行验证性测试.结果表明,MELM 获得理想的微信热点预测结果,而且微信热点预测的实时性良好.

2 改进极限学习机

2.1 标准极限学习机

对一个具体的预测问题,首先要采集历史数据,设历史数据组成的样本集为 $\{(x_j, t_j)\}_{j=1}^k, x_j = [x_j, \dots, x_{j+n-1}]^T, t_j = x_{j+n}, x_j, n$ 为数据的特征维数; k 为样本数量,训练样本的回归误差为 ϵ_i ,那么标准极限学习机的回归形式可以表示为

$$\min \left(\frac{1}{2} \beta_L^T \beta_L + \frac{\gamma}{2} \epsilon^T \epsilon \right) \quad (1)$$

s. t.

$$t_j = \sum_{i=1}^L \beta_i f(\alpha_i x_j + c_i) - \epsilon_j$$

式中, L 表示隐含层节点的数; c_i 表示第 i 个隐含层节点的误差; $f(\cdot)$ 表示隐含层节点的非线映射函数; α_i 和 β_i 分别表示隐含层节点和输入节点的权值.

对式(1)进行求解,效率相当低而且难以获得全局最优解,为此引拉格朗日乘子建立如下优化函数:

$$L(\omega, \epsilon, \beta_L) = \frac{1}{2} \beta_L^T \beta_L + \frac{\gamma}{2} \epsilon^T \epsilon - \omega (\mathbf{H}_L \epsilon_L - T - \epsilon) \quad (2)$$

式中, $\mathbf{H}_L$ 表示节点矩阵; ω 表示权值; T 表示输出结果.

对变量的偏导进行计算,得到如下公式

$$\begin{cases} \frac{\partial L}{\partial \beta_L} \rightarrow \beta_L^T = \omega \mathbf{H}_L \\ \frac{\partial L}{\partial \epsilon} \rightarrow \gamma \epsilon^T + \omega = 0 \\ \frac{\partial L}{\partial \omega} \rightarrow \mathbf{H}_L \beta_L - T - \epsilon = 0 \end{cases} \quad (3)$$

对式(3) 求解,极限学习机的权值计算公式为

$$\beta_L = (\mathbf{H}_L^T \mathbf{H}_L + \frac{\mathbf{I}_L}{\gamma})^{-1} \mathbf{H}_L^T T \quad (4)$$

利用权值计 β_L 构建相应问题的预测模型,得到输出结果的表达式为

$$t = \sum_{i=1}^L \beta_i f(\alpha_i x + c_i) \quad (5)$$

2.2 改进极限学习机(MELM)

在标准极限机学习过程中,关键要找到 β_L 的最优值,在 β_L 求解过程,有大量的矩阵求逆运算,导致计算复杂度高,对微信热点预测模型的训练过程产生不利影响,因此本文对进行相应的改进,引入 Cholesky 分解方法加极限学习机的训练过程,快速找到 β_L 的最优求解.根据式(3) 能够得到

$$\begin{cases} \beta_L = \omega \mathbf{H}_L^T \\ \epsilon = -\frac{\omega T}{\gamma} \end{cases} \quad (6)$$

利用式(6) 对 β_L 进行求解,可以得到相应的线性方程形式为

$$\mathbf{A}_L \beta_L = \mathbf{b}_L \quad (7)$$

同时满足如下约束条件:

$$\begin{cases} \mathbf{A}_L = \mathbf{H}_L^T \mathbf{H}_L + \frac{\mathbf{I}_L}{\gamma} \\ \mathbf{b}_L = \mathbf{H}_L^T T \end{cases} \quad (8)$$

综合式(6) 和式(8) 可以得到

$$\mathbf{A}_L^T = \left(\mathbf{H}_L^T \mathbf{H}_L + \frac{\mathbf{I}}{\gamma} \right)^T = \mathbf{H}_L^T \mathbf{H}_L + \frac{\mathbf{I}}{\gamma} = \mathbf{A}_L \quad (9)$$

对于 $\mathbf{v}, \mathbf{A}_L$ 的二次型可以描述为

$$\begin{aligned} \mathbf{v}^T \mathbf{A}_L \mathbf{v} &= \mathbf{v}^T \left(\mathbf{H}_L^T \mathbf{H}_L + \frac{\mathbf{I}}{\gamma} \right) \mathbf{v} \\ &= \sum_{i=1}^L (f(\alpha_i x_1 + c_i) v_i)^2 + \dots + \\ &\quad \sum_{i=1}^L (f(\alpha_i x_k + c_i) v_i)^2 + \frac{1}{\gamma} \sum_{i=1}^L v_i^2 \end{aligned} \quad (10)$$

对式(10) 进行详细分析可以发现, $\mathbf{A}_L$ 是一个对称正定矩阵,那么采用 Cholesky 对其进行分解,可以得到:

$$\mathbf{A}_L = \mathbf{S}_L \mathbf{S}_L^T \quad (11)$$

式中, $\mathbf{S}_L$ 是一个三角矩阵.

s_{ij} 表示三角矩阵 $\mathbf{S}_L$ 中的非零元素,那么根据 $\mathbf{A}_L$ 的元素 α_{ij} 可以得到

$$s_{ij} = \begin{cases} \sqrt{a_{ii} - \sum_{n=1}^{i-1} s_{in}^2}, i = j \\ (a_{ij} - \sum_{n=1}^{j-1} s_{in}s_{jn})/s_{jj}, i > j \end{cases} \quad (12)$$

式中, $i = 1, 2, \dots, L; j = 1, 2, \dots, L$.

结合式(11)和式(7),同时乘上 $\mathbf{S}_L^{-1}$,那么根据 $\mathbf{F}_L = \mathbf{S}_L^{-1}\mathbf{b}_L$ 与 $\mathbf{S}_L\mathbf{F}_L = \mathbf{b}_L$ 等价,可以得到 $\mathbf{F}_L$ 的元素 f_i 为

$$f_i = \begin{cases} b_i/s_{ii}, i = 1 \\ (b_i - \sum_{n=1}^{i-1} s_{in}f_n)/s_{ii}, i > 1 \end{cases} \quad (13)$$

式中, $i = 1, 2, \dots, L$.

根据 $\mathbf{S}_L$ 和 $\mathbf{F}_L$ 可以得到 β_L 的计算公式为

$$\beta_i = \begin{cases} f_i/s_{ii}, i = L \\ (f_i - \sum_{n=1}^{L-i} s_{i+n,i}\beta_{i+n})/s_{ii}, i < L \end{cases} \quad (14)$$

对比 ELM 和 MELM 的学建模过程,MIELM 的 β_L 求解仅通过四则运算就可以实现,没有矩阵求逆运算,计算更加简单,大幅度减少了求解的时间。

尤其当隐含层的节点增多条件下,速度加快,可以得到:

$$\mathbf{H}_{L+1} = \begin{bmatrix} \mathbf{H}_L \\ \mathbf{h}_{L+1} \end{bmatrix} = \begin{bmatrix} \mathbf{h} \\ \mathbf{h}_L \\ \mathbf{h}_{L+1} \end{bmatrix} \quad (15)$$

那么, $\mathbf{A}_{L+1}$ 与 $\mathbf{A}_L$ 之间的关系可以表示为

$$\mathbf{A}_{L+1} = \mathbf{H}_{L+1}^T \mathbf{H}_{L+1} + \frac{\mathbf{I}_{L+1}}{\gamma} = \begin{bmatrix} \mathbf{A}_L & \mathbf{Q}_{L+1} \\ \mathbf{Q}_{L+1}^T & \mathbf{P}_{L+1} \end{bmatrix} \quad (16)$$

式中, $\mathbf{Q}_{L+1} = [\mathbf{h}_1^T \mathbf{h}_{L+1} \dots \mathbf{h}_L^T \mathbf{h}_{L+1}]^T$, $\mathbf{P}_{L+1} = \mathbf{h}_{L+1}^T \mathbf{h}_{L+1} + \frac{1}{\gamma}$.

根据 Cholesky 分解过程可以知道,通过计算 $\mathbf{s}_{L+1,1}$ 与 $\mathbf{s}_{L+1,L}$ 中不为零的元素就可以得到 $\mathbf{S}_{N+1}$,此时可以得到

$$\mathbf{b}_{L+1} = \mathbf{H}_{L+1}^T \mathbf{T} \begin{bmatrix} \mathbf{b}_L \\ \mathbf{h}_L^T \mathbf{T} \end{bmatrix} \quad (17)$$

那么,根据 $\mathbf{F}_L$ 可以得到

$$\mathbf{F}_{L+1} = \begin{bmatrix} \mathbf{F}_L \\ \mathbf{f}_{L+1} \end{bmatrix} \quad (18)$$

因此根据 $\mathbf{f}_{L+1}$ 可以得到 $\mathbf{F}_{L+1}$,不要重新计算 $f_1, f_2, \dots, f_L$,加快了学习效率,而且可以实现 ELM 的在线学习。

3 MELM 的微信热点预测模型

3.1 微信热点数据的混沌分析

微信热点通常是一个一维的数据: $\{x(i), i = 1, 2, \dots, n\}$,根据其混沌性得到延迟时间(τ)和嵌入维(m),然后把它变化一个多维数据: $X(t) = \{x(t), x(i+\tau), \dots, x(i+(m-1)\tau)\}$,从而将表面上看起来没有规律的数据变为实际有规律的数据,从而发现其中包含的变化特点.采用关联积分算法确定最优的 τ 和 m .设两个样本点为 $X(i)$ 和 $X(j)$,它们的距离 $r_{ij}(m)$ 为

$$r_{ij}(m) = \|X(i) - X(j)\| \quad (19)$$

那么可以得到关联积分为

$$C(m, N, r, \tau) = \frac{2}{M(M-1)} \sum_{1 \leq i < j \leq M} H(r - \|X(i) - X(j)\|) \quad (20)$$

式中, r 为距离的阈值。

全部样本划分为 t 个序列, C_l 为第 l 个序列的相关积分,那么可以得到:

$$S(m, r, \tau) = \frac{1}{t} \sum_{l=1}^t \{C_l(m, r, \tau) - C_l(l, r, \tau)^m\} \quad (21)$$

可以得到极小值点为

$$\Delta \bar{S}(t) = \frac{1}{4} \sum_{m=2}^5 \max[S(m, r_j, t)] - \min[S(m, r_j, t)] \quad (22)$$

当 $\Delta \bar{S}(t)$ 达到最小值时,那么可以认为得到最优的 τ .

变换后第 i 个向量为 $X_i(m+1)$,最近邻为 $X_{n(i,m)}(m+1)$,则有

$$\alpha(i, m) = \frac{\|X_i(m+1) - X_{n(i,m)}(m+1)\|}{\|X_i(m) - X_{n(i,m)}(m)\|} \quad (23)$$

设 $E(m) = \frac{1}{N - m\tau} \sum_{i,j=1}^{N-m\tau} \alpha(i, m)$,若 $E(m)$ 处于最大值状态时,那么就可以认为得到最优的 m .

3.2 MELM 的微信热点预测步骤

(1) 对人们关心的一个微信热点话题进行分析,采集其历史点击率,构成微信热点话题的学习样本。

(2) 采用关联积分算法估计 τ 和 m ,并对微信热点话题数据进行变换,得到训练样本和测试样本。

(3) 采用改进极限学习机训练微信热点话题样

本,建立微信热点话题的预测模型.

(4)对测试样本进行预测,根据预测结果对模型性能进行分析.

4 模型性能的测试

4.1 样本的构建

选择“天津化工厂爆炸”这个微信热点作为研究对象,采集的数据如图 1 所示. 对图 1 进行分析,可以发现该微信热点变化很复杂,不仅具有一定的增长趋势,同时具有强烈的波动性. 采用关联积分算法估计 τ 和 m ,结果如图 2 和图 3 所示,从图 2 和图 3 可知,最优的 τ 和 m 分别为:7 和 8,根据 $\tau=7$ 和 $m=8$ 得到微信热点的多维数据,选择前 200 个数据作为训练样本,其余用于测试其预测效果.

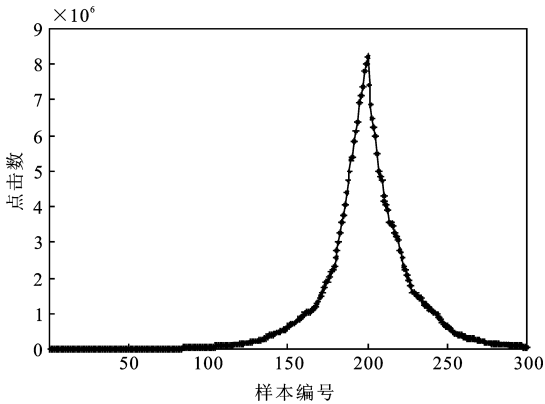


图 1 微信热点的数据

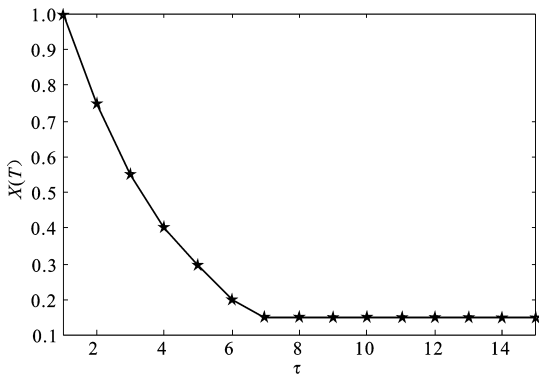


图 2 估计最优的 τ

4.2 结果与分析

采用 MELM 对天津化工厂爆炸”这个微信热点话题进行预测,结果如图 4 所示. 对图 4 进行分析可以发现,MELM 的微信热点测精度高,而且预测结果十分稳定,表明 MELM 可以用于微信热点话题预测中,而且预测结果十分理想.

选择当前经典模型:多元线性回归(MLR)、BP

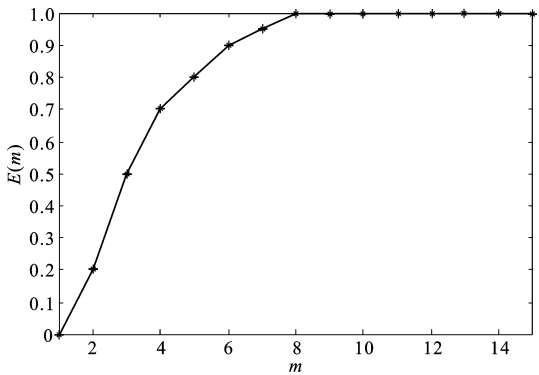


图 3 估计最优的 m

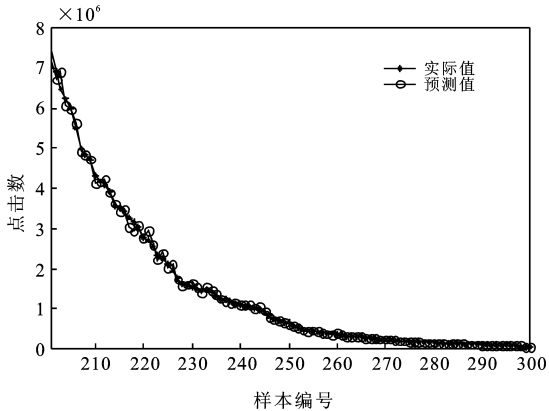


图 4 MELM 的预测效果

神经网络(BPNN)、支持向量机(SVM)对 2016 年的 10 个微信热点进行预测,它们预测结果如表 1 所示. 对比它们微信热点的预测精度可以发现,相对于当前经典模型,MELM 的微信热点预测精度有了一定的提高,而且预测结果更优,这表明,MELM 可以很好地对微信热点进行建模,把握其变化态趋,是一种通用性强的微信热点预测模型.

表 1 预测精度统计					%
热点话题	MLR	BPNN	SVM	MELM	
南海仲裁事件	83.02	89.02	91.19	95.34	
美国大选	83.19	91.70	94.96	96.68	
魏则西事件	83.44	88.23	94.32	96.43	
女排夺冠	80.50	88.78	91.66	96.07	
A 股熔断事件	80.86	88.05	91.64	96.41	
朴景惠“闺蜜案”	80.08	89.83	94.44	96.82	
马宝强离婚	82.99	90.13	93.96	96.92	
雷洋事件	80.79	90.62	92.54	96.38	
G20 峰会	84.57	89.39	91.68	96.20	
连云港反核群体事件	83.13	90.60	91.25	95.05	

统计所有模型的平均建模时间,结果如表 2 所

示,可以发现,MELM 的平均建模时间最少,加快了微信热点的建模效率,可以实现微信热点在线预测.

表 2 平均建模时间统计

热点话题	MLR	BPNN	SVM	MELM
南海仲裁事件	10.36	6.42	20.05	3.83
美国大选	11.68	5.38	20.26	3.25
魏则西事件	10.25	6.00	20.97	3.84
女排夺冠	11.50	5.56	22.80	3.68
A 股熔断事件	11.85	5.65	20.62	4.38
朴景惠“闺蜜案”	11.16	5.03	20.19	3.53
马宝强离婚	10.67	6.19	24.51	3.38
雷洋事件	10.59	6.66	21.47	4.24
G20 峰会	10.17	6.67	22.92	4.24
连云港反核群体事件	10.42	6.90	21.18	3.46

5 结束语

微信热点受到人的思想、政治、经济以及其他因素的影响,变化十分复杂,不仅具有强烈的时变性,而且具有一定的混沌变化特点,针对当前微信热点预测模型的局限性,以获得更加理想的微信热点预测结果,设计了基于 MELM 的微信热点预测模型,采用大量的仿真实验测试与当前经典模型的优越性,结果表明,MELM 的微信热点预测效果更佳,预测结果更加稳定、可信,而且建模预测效率更高,可以应用于实际网络舆情数据分析,预测结果可以帮助预控一些负面微信热点扩散,具有较高的实际应用价值.

参考文献:

[1] 路斌. 互联网舆情热点自动发现与分析技术研究与实践[D]. 北京:北京大学,2007.

[2] 戴臻. 布谷鸟算法优化支持向量机的网络热点话题预测[J]. 计算机应用与软件, 2014, 31(4): 330-333.

[3] 李睿, 赵保华, 王朋. 基于遗传优化加权相关向量机的网络热点预测模型[J]. 计算机应用与软件, 2015, 32(6): 312-316.

[4] 刘玉坤,夏栋梁,马 丽. 基于 AGSO-LSSVM 的热点话题预测模型 [J]. 重庆邮电大学学报(自然科学版), 2014, 26(6): 803-808.

[5] 蒋玉婷. 支持向量机修正 ARIMA 误差的微博热点预测[J]. 计算机应用与软件, 2014, 31(9): 187-190.

[6] 梁野,郭宁宁,李碧菽,等. 基于机器学习的网络媒体热点话题预测方法研究与实现[J]. 微型机与应用, 2014, 33(15): 62-64.

[7] 罗泰晔. 网络舆情热点规模预测研究[J]. 情报杂志, 2016, 35(10): 181-185.

[8] 张乐,祁超. 网络论坛热点话题的关注度预测[J]. 计算机与数字工程,2013,41(5):772-775.

[9] 卢珥珈,张宏莉,张玥. 基于 BBS 的热点话题发现与态势预测技术的研究[J]. 智能计算机与应用,2012,2(2):1-5.

[10] 陈江,刘玮,巢文涵,等. 融合热点话题的微博转发预测研究[J]. 中文信息学报,2015,29(6):150-158.

[11] 张弦,王宏力. 基于 Cholesky 分解的增量式 ELM 及其在时间序列预测中的应用[J]. 物理学报,2011,60(11):201-207.

[12] 李郁侠, 赵军科,王伟,等. 电价时间序列混沌特性分析及电价预测[J]. 西北农林科技大学学报(自然科学版),2008,36(1):229-234.

作者简介:

林 荫 女,(1970-),硕士,讲师. 研究方向为算法分析与设计、数据处理技术 . E-mail:linyinsj@sina. com.

杨长春 男,(1963-),博士,教授. 研究方向为数据挖掘、大数据应用技术.