基于收视行为的互联网电视节目流行度预测模型

朱琛刚 程 光*
(东南大学计算机科学与工程学院 南京 211189)
(教育部计算机网络和信息集成重点实验室(东南大学) 南京 211189)

摘 要：准确预测节目流行度是互联网电视节目系统设计与优化所要解决的关键问题之一。针对现有预测方法存在模型训练时间长、样本数量多、且对突发热点节目流行度预测效果差等问题，该文测量了某互联网电视平台 280 万用户的 60 亿条收视行为数据，采用行为动力学分类方法将节目流行度演化过程分为内源临界、内源亚临界、外源临界和外源亚临界 4 种类型，运用双种群粒子优化的最小二乘支持向量机对每种类型分别构建了一种互联网电视节目流行度预测模型 BD3P，并将 BD3P 模型应用于实际数据测验。实验结果表明，与现有其他方法相比，BD3P 模型预测精度可提升 17%以上，并能有效缩短预测周期。

关键词：互联网电视；流行度预测；行为动力学；最小二乘支持向量机；双种群粒子群优化
中图分类号： TP393 **文献标识码**： A **文章编号**：1009-5896(2017)10-2504-09
DOI: 10.11999/JEIT161310

Program Popularity Prediction Model of Internet TV Based on Viewing Behavior

ZHU Chengang CHENG Guang
(*School of Computer Science and Engineering*, *Southeast University*, *Nanjing* 211189, *China*)
(*Key Laboratory of Computer Network and Information Integration*, *Ministry of Education* (*Southeast University*), *Nanjing* 211189, *China*)

Abstract: Predicting program popularity is a key issue for design and optimization of Internet TV system. Existing prediction methods usually need large quantity of samples and long training time, while the prediction accuracy is poor for the burst hot programs. This paper introduces an Internet TV Program Popularity Prediction model based on viewing Behavioral Dynamics features (BD3P). 6 billion view behavior records from 2.8 million subscribers of a certain Internet TV platform are measured, and the evolution process of program popularity is divided into 4 types based on behavioral dynamics features, which is endogenous, internal subcritical, exogenous and exogenous subcritical. The prediction models of Internet TV program popularity are constructed for each type using Least Squares Support Vector Machines (LSSVM) with double population Particle Swarm Optimization (PSO), and these models are applied to the actual data test. The experimental results show that, compared to the existing prediction model, the prediction accuracy can be increased by more than 17%, and the forecast period can be effectively shortened.

Key words: Internet TV; Popularity prediction; Behavioral Dynamics (BD); Least Squares Support Vector Machines (LSSVM); Double population Particle Swarm Optimization (PSO)

1 引言

互联网电视顺应了高清化、互动化、融合化、智能化的发展趋势，将“按需点播”的原则应用到电视业务中，在向用户提供直播服务的同时，允许用户采用暂停、快进、快退及点播等方式收看特定时间或栏目的内容，成为“三网融合”的代表性业务。根据第 38 次《中国互联网络发展状况统计报告》[1]，截至 2016 年 6 月，我国网络视频用户规模达 5.14 亿，较 2013 年底增加 8400 万；网络直播用户规模达到 3.25 亿。互联网电视的流量从 2013 年每月 1.3 EB，达到 2016 年每月 6.5 EB，增长了 5 倍。然而用户有限的注意力往往被少数视频所吸引。

收稿日期：2016-12-08；改回日期：2017-06-15；网络出版：2017-07-21
*通信作者：程光 gcheng@njnet.edu.cn
基金项目：国家 863 计划项目(2015AA015603)，江苏省未来网络创新研究院未来网络前瞻性研究项目(BY2013095-5-03)，江苏省“六大人才高峰”高层次人才项目(2011-DZ024)
Foundation Items: The National 863 Program of China (2015AA 015603), The Prospective Research Program on Future Networks of Jiangsu Province (BY2013095-5-03), The Six Industries Talent Peaks Plan of Jiangsu Province (2011-DZ024)

根据腾讯视频的电视剧总播放排行榜[2]，累计收视量排名前三的连续剧收视量都超过 45 亿次，收视量前 50 名的电视连续剧最低都在 10.7 亿次以上，远超过其他电视节目的收视量。

在此情况下，预测互联网电视节目的流行度变得十分必要。首先，通过在推荐系统引入节目流行度预测结果，互联网用户可以便捷地从海量视频资源中定位到更有价值的内容。其次，商业公司基于节目流行度预测数据，调整广告投放策略，可以最大化广告的宣传效果，带来丰厚的投资回报。再者，网络运营商可以采用流行度预测模型优化现有的视频覆盖网络，根据未来用户的需求提前调配网络中的传输和存储资源用于热点节目的分发，从而避免网络拥塞，提升用户体验，扩大市场占有率。文献[3,4]中采用内容流行度作为信息中心网络缓存调度策略的测算依据，有效地改善了网络访问性能。然而，预测电视节目的流行度是一项极具挑战性的工作。这是由于内容质量、节目内容与收视用户之间的关联性等一系列影响节目流行度的关键因素是难以量化测量的；如何捕捉节目内容与现实世界中事件之间的关系，并将其引入到预测模型中也是颇具难度的问题；受到节目收视观众群的影响，不同节目的流行度随时间变化差异巨大，预测模型必须具备自适应动态调整的能力。现有的流行度预测方法都是面向其他媒体形式，但可以作为借鉴。预测方法主要有累计增长法[5−7] (cumulative growth)、时间序列分析法[8−10] (temporal analysis)和演化趋势分析法[11−13] (evolution trends)，采用的数据源主要分为聚合行为和个体行为[14]。传统方法更多考虑的是将个体行为看成一个随机过程，充分利用已有的数据训练集来预测将来某时刻的流行度与已有数据之间的关系，不区分可能的不同流行度发展行为。而本文所提出的动力学模型是以 Crane 等人[10]所提出的理论为基础，该理论明确指出个体的随机观看行为与个体之间的相互作用结构共同影响了流行度的发展。这与以往单纯基于个体行为以及相关时间序列的分析是不同的，结果大幅提高了模型的预测精度。

本文从分析节目流行度时间序列的演化规律入手，基于社交行为动力学理论归纳出收视行为的动力学特征，提出一种互联网电视节目流行度预测模型 BD3P(Program Popularity Prediction model based on viewing Behavioral Dynamics features)，并采用粒子群算法优化模型的寻优性能。首先综合考虑流行的起源和节目传播过程，将互联网电视平台节目的流行度演化趋势归纳为 4 类；然后为减轻不同节目时间序列数据之间的相互干扰，运用最小二乘支持向量机(Least Squares Support Vector Machines, LSSVM)对 4 种演化趋势分别进行建模；接着为避免人为选择函数的盲目性，使用同时具有全局最优与局部最优的双种群粒子群算法优化 LSSVM 的超参数；最后将 4 种模型的预测结果按分类概率叠加得到节目流行度的预测值。本文创新点主要包括以下两个方面：(1)采用行为动力学分类方法对某广电运营商互联网电视平台 210 天 11 万条节目约 60 亿条的收视记录进行分析，综合考虑节目流行起源和传播过程，将节目流行度的演化过程分为内源亚临界、内源临界、外源亚临界、外援临界 4 种类型。(2)基于流行度演化过程分类，提出一种新的互联网电视节目流行度预测模型 BD3P。采用双种群粒子群算法优化的 LSSVM 算法，减小了预测误差，缩短了观测周期。以预测节目上线 30 天内的流行度为例，BD3P 模型的预测均方根误差低于 0.17，较基准模型误差减少 17.1%。同时所需训练样本数量也较少，仅需多元线性模型(Multivariate Linear, ML)75%的数据量，就可以使得预测结果与真实值之间的复相关系数达到 95%。

本文第 2 节给出了流行度预测问题的形式化定义，根据行为动力学理论分析了互联网电视平台上节目流行度变化趋势，并采用双种群粒子群优化 LSSVM 算法构建流行度预测模型；第 3 节，基于某广电运营商互联网电视平台 280 万用户 213 天的收视数据，采用均方根误差和复相关系数(R^2)两个评估指标将 BD3P 与 3 种基准算法进行对比，以评估模型效果；第 4 节总结全文并展望未来的工作。

2　BD3P 预测模型

2.1 问题定义

本研究工作中对互联网电视节目流行度定义为：给定一个节目 c，以节目播出时间为 0 时刻，记在 t 时刻的收看次数为 $N(t)$，即节目在 t 时刻的流行度。节目流行度预测模型的目标是根据预测节目 c 上线 t_i 天的收视次数，预测其在 t_r 天时的收视次数($t_i < t_r$)。设 $c \in C$ 代表节目集 C 在播出时间段 T 中播出的某部节目，$t \in T$ 表示节目已经上线的时长。观测时间 t_i 表示对节目流行度做出预测的时间点。t_i 越大，则已知的流行度数据就越多。t_r 表示需要预测节目流行度的目标时间点。设 $N_c(t_i)$ 表示节目 c 从上线到做出预测时的流行度，$N_c(t_r)$ 表示该节目稍后在 t_r 时刻的流行度，$\widehat{N}_c(t_i,t_r)$ 表示使用节目 c 在 t_i 时刻信息计算出的其在 t_r 时刻流行度的预测值。$\widehat{N}_c(t_i,t_r)$ 与 $N_c(t_r)$ 越接近，表示预测模型越有效。

2.2 收视行为的动力学特征

根据社交行为背后的动力学原因，本文此处用两个要素来构建解释互联网电视节目观看行为的收视行为动力学模型。第 1 个要素是用来描述人类行为的幂律型响应函数，它刻画了以上这些原因的潜在影响(式(1))。由定义，记忆核函数 $\phi(t)$ 表示对于单个个体而言，从“原因”到“行动”(观看视频)的等待时间分布。“原因”可能是上述的任何一种，而“行动”指的是在某种单一原因的直接影响下(不考虑其它原因)经过时间 t 后观看了视频。

$$\phi(t) \approx 1/t^{1+\theta},\ \ 0<\theta<1 \tag{1}$$

第 2 个要素是传播的分叉过程，考虑社交网络上的层级影响。特别是对于连通性比较高的网络，一个“原因”可以通过中间步骤引发多个“行动”。比如，在对视频有相似兴趣的朋友圈中，视频观看行为可以由用户甲传给用户乙，用户乙传给用户丙这么无限传播下去。这种级联的传播过程可以由自激 Hawkes 条件 Poisson 过程描述(式(2))。$\lambda(t)$ 表示视频的瞬时观看次数。

$$\lambda(t) = V(t) + \sum_{i,t_i \le t} \mu_i \phi(t-t_i) \tag{2}$$

其中，μ_i 指的是第 i 个人可能影响到的在 $[t_i,t]$ 这段时间内去观看视频的人数。如果社交网络的连通性好，即每个人可能影响到的人很多，则 μ_i 的值可以很大。$V(t)$ 是外源的，它描述了自发的观看行为，可以理解为 t 时刻有 $V(t)$ 个人正好看到了某个视频，而没人向这些人推荐。

基于上述分析，本文建立了互联网电视节目收视的行为动力学分类模型。根据扰动类型(内/外)，以及个体影响他人“行动”的能力(临界/亚临界)，即在式(1)与式(2)的相互作用下，用户收视节目的动力学行为对节目流行度演化的影响可以分成 4 种类型，并可以由一个共同参数的 θ 联系起来。图 1 描述了互联网电视节目流行度的 4 种演化模型。

(1)内源亚临界类：此时观看行为是随机的，流行度变化趋势如图 1(a)所示，观看行为可用式(3)描述。$\eta(t)$ 是一个描述收视量随机变化的函数。

$$\widehat{N}_c(t_i,t_r) = \eta(t) \tag{3}$$

(2)内源临界类：这种情况下不但社交网络成熟，而且收视行为由于网络内部的互相推荐而得到增强，流行度变化趋势如图 1(b)所示。收视量变化的特征为随时间按 $1\text{–}2\theta$ 的幂律关系上升到峰值后再按近似速率下降，收视行为可用式(4)描述。

$$\widehat{N}_c(t_i,t_r) \approx \frac{1}{\left|t_r - t_i\right|^{1-2\theta}} \tag{4}$$

(3)外源亚临界类。这种情况下社交网络不成熟，连通性较差，平均的 μ_i 小于 1，这样在 t_i 时刻的一个外源性事件诱发的活动不会传出很多代，此时只与记忆函数直接相关，流行度变化趋势如图 1(c)所示。收视量变化特征为快速达到峰值后随时间按 $1+\theta$ 的幂律关系快速衰减，收视行为可用式(5)描述。

$$\widehat{N}_c(t_i,t_r) \approx \frac{1}{(t_r - t_i)^{1+\theta}} \tag{5}$$

(4)外源临界类。此时社交网络较为成熟，即平

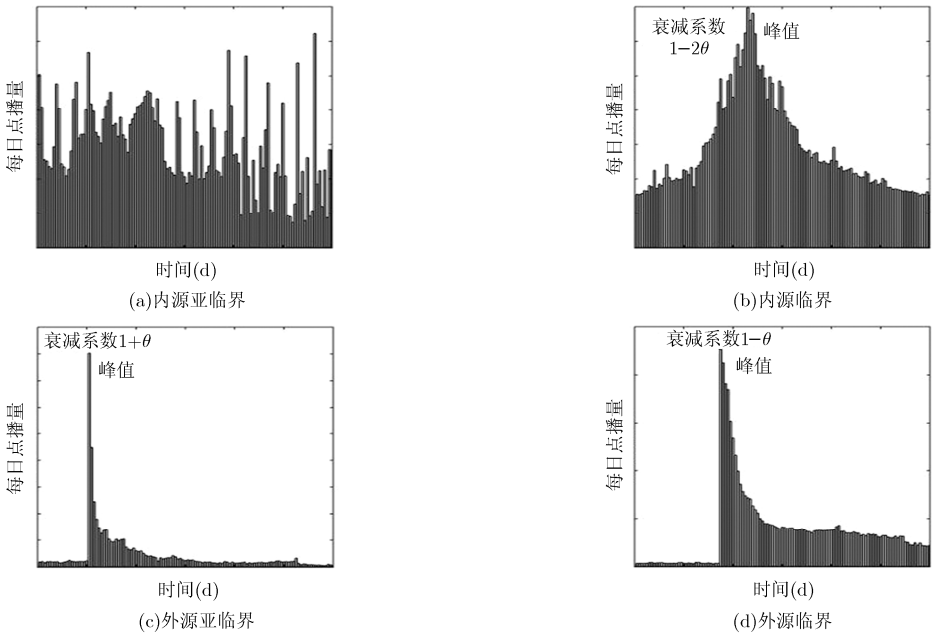


图 1 互联网电视节目流行度演化模型分类

均的 μ_i约等于 1，外源性事件诱发的收视行为可以传出很多代，流行度变化趋势如图 1(d)所示。收视量变化特征为快速达到峰值后按随时间按 $1-\theta$ 的幂律关系缓慢衰减，收视行为可用式(6)描述。

$$\widehat{N}_c(t_i,t_r)\approx\frac{1}{(t_r-t_i)^{1-\theta}} \tag{6}$$

幂律响应函数最重要的参数是θ，而传播分叉过程最重要的参数是 μ_i 的期望值 n（定义为$n\equiv\langle\mu_i\rangle$）。$\theta$ 和 n 对于模型的理论分析结果具有决定性的影响。而其他参数，比如记忆核函数$\phi(t)$具体表达形式中的参数，则对理论结果没有定性影响。θ (0 $<\theta<$1)和 n 决定了上述 4 种流行度发展行为，考虑更为细节的其他参数选择并不会导致更多类型的流行度模式。关键参数幂指数θ可以通过拟合节目集收视量和时间对数关系的斜率得到，如图 2 所示。

本文研究了某广电运营商互联网电视平台 2016 年 1 月 1 日至 2016 年 7 月 31 日共计 213 天 280 万用户的收视记录，得到了 4 种类型共计 11 万条节目流行度演化趋势，计算出幂指数θ=0.4±0.1。这与文献[10]在流行度趋势分类的系统定性上是相符的。

2.3 最小二乘支持向量机预测模型

自回归模型(Auto-Regressive and Moving Average model, ARMA)等经典的多维时间序列分析模型属于线性模型，实际预测能力较弱。神经网络算法虽然具有较好的非线性逼近能力，但是容易陷入局部最优，而且可解释性差，需要丰富的建模经验予以辅助。支持向量机(Support Vector Machine, SVM)遵循结构风险最小的统计学习理论，较好地解决了局部最优、过拟合、非线性等问题，且具有较好的泛化能力。LSSVM 将 SVM 的求解从二次规划问题转化为线性方程组，加速了训练过程。由于径向基函数(Radial Basis Function, RBF)可以将特征向量映射到一个非线性空间，易于捕捉到早期和后期收视量之间的相关性。因此本文选择 LSSVM 作为流行度预测的回归模型，采用 RBF 作为 LSSVM 的核函数，节目 c 流行度的预测值 $\widehat{N}_c(t_i,t_r)$ 可以用式(7)、式(8)确定：

$$\widehat{N}_c(t_i,t_r)=\sum_{k=1}^{K}\alpha_k\cdot\Phi(\boldsymbol{X}(c,t_i),\boldsymbol{X}(k,t_i))+b \tag{7}$$

$$\Phi(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{8}$$

其中，$\Phi(x,y)$是高斯径向基核函数；σ 为可调参数，表示核函数的宽度；$\boldsymbol{X}(c,t_i)$是节目 c 在t_i时刻的时间序列向量。系数α_k和截距 b 为回归因子，可以通过最小化正则风险泛函式(9)获得α_k和 b：

$$R(\omega)=\frac{1}{2}\|\omega\|^2+D\sum_{i=1}^{m}L(f(x_i),y_i) \tag{9}$$

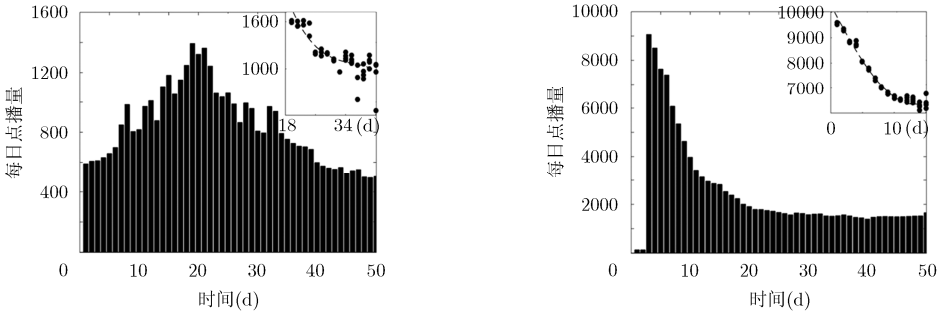
式(9)中，常数 D 为惩罚系数，$L(f(x_i),y_i)$表示损失函数，通常采用一次不敏感函数(式(10))。

$$L_\varepsilon(f(x),y)=\max\{|f(x)-y|-\varepsilon,0\} \tag{10}$$

惩罚参数 D、核函数参数σ 以及相关损失函数参数ε 是影响 LSSVM 回归性能的 3 个超参数。为避免人为选择函数的盲目性，提高预测模型准确度，需要优化模型参数。本文采用双种群粒子群优化算法对模型参数进行优化。

2.4 粒子群算法优化 LSSVM 超参数

粒子群优化算法(Particle Swarm Optimization, PSO)是一种通过模拟自然界的生物活动以及群智能的随机全局搜索算法。PSO 算法每次迭代过程中，粒子都依靠全局最优值来更新，容易造成粒子多样性消失，趋于统一化，从而造成算法收敛的速度较慢。为解决上述缺陷，Wang 等人[15]提出了具有局部寻优和全局寻优的双种群粒子群优化算法(Double Population PSO, DP-PSO)，采用动态调整加速因子的方法提高算法的寻优性能。具体方法为：假设粒子群由 s 个粒子组成，把它分成两个群体 Q_1(负责局部寻优)和 Q_2(负责全局寻优)，Q_1 由 s_1 个粒子组成，Q_2 由 s_2 个粒子组成(s=s_1+s_2)。两个群体采用不同的进化过程，种群 Q_1 采用快速收敛



(a)内源临界类幂律响应函数的参数θ　　(b)外源临界类幂律响应函数的参数θ

图 2 不同演化类型的幂律响应函数的参数 θ

的进化方程，以增强局部寻优能力；种群 Q_2 采用全局搜索能力的进化方程，其中加速因子采用反正切调整策略。进化方程如式(11)～式(13)：

$$\begin{aligned} Q_1: v_{ij}(t+1) &= \omega \times v_{ij}(t) + c_1 \times \text{rand}() \times (p_{ij}(t) \\ &\quad - x_{ij}(t)) + c_2 \times \text{rand}() \\ &\quad \times (p_{gj}(t) - x_{ij}(t)) \end{aligned} \tag{11}$$

$$\begin{aligned} Q_2: v_{ij}(t+1) &= \omega \times v_{ij}(t) + c_1 \times r_{1j}(t) \\ &\quad \times (p_{ij}(t) - x_{ij}(t)) + c_2 \times r_{2j}(t) \\ &\quad \times (p_{gj}(t) - x_{ij}(t)) \end{aligned} \tag{12}$$

$$\omega(t) = 0.9 - \frac{t}{t_{\max}} \times 0.5 \tag{13}$$

加速因子 c_1 和 c_2 代表将每个粒子推向局部最优和全局最优位置的统计加速项的权重。DP-PSO 算法在搜索初期要使粒子尽可能地飞跃整个搜索空间，以获得粒子的多样性；在搜索末期，使粒子以较快的速度，精确收敛于全局最优解。通过分析加速因子的变化对算法影响，利用反正切函数动态调整 c_1 和 c_2 的策略，更好地平衡全局搜索和局部搜索。c_1，c_2 的取值公式分别为

$$\begin{aligned} c_1(t) &= \big(c_{1\mathrm{s}} - (c_{1\mathrm{s}} - c_{1\mathrm{e}}) \times \arctan\big(20 \times t / T_{\max} - e\big) \\ &\quad + \arctan e\big) \big/ l \end{aligned} \tag{14}$$

$$\begin{aligned} c_2(t) &= \big(c_{2\mathrm{s}} - (c_{2\mathrm{s}} - c_{2\mathrm{e}}) \times \arctan\big(20 \times t / T_{\max} - e\big) \\ &\quad + \arctan e\big) \big/ l \end{aligned} \tag{15}$$

其中，$c_{1\mathrm{s}}$ 和 $c_{2\mathrm{s}}$ 分别是 c_1 和 c_2 的初值；$c_{1\mathrm{e}}$ 和 $c_{2\mathrm{e}}$ 分别是 c_1 和 c_2 的终值；$T_{\max}$ 为算法的最大迭代次数；e 为调节系数，控制曲线的衰减，一般取值为 0～10；$l = \arctan(20 - e) + \arctan e$。

基于 DP-PSO 的 SVM 参数优化流程图如图 3 所示，具体选择步骤描述如下：

步骤 1　初始化粒子群 (D, σ, ε)、惯性权重 ω，加速因子 c_1，c_2，种群个数以及最大迭代次数等。

步骤 2　使用支持向量机对训练样本进行回归训练，得出每个粒子的适应度值，记录粒子的个体最优位置和全局最优位置。

步骤 3　根据式(11)～式(15)对位置和速度进行更新。

步骤 4　使用支持向量机对训练样本进行回归训练，得出每个粒子的适应值。将每个粒子当前位置及种群中的所有粒子所经历的最好位置进行比较，如果这个粒子的位置较优，则将其设置为当前的最好位置；否则，最好位置保持不变。

步骤 5　返回步骤 3，直到满足最大迭代次数 $T_{\max}$ 或达到要求的误差，则终止迭代，输出此时的全局最优位置，用得到的最优位置对测试样本进行回归预测。

2.5 建模流程

综合考虑了 4 种节目流行度演化类型，BD3P 模型采用双种群粒子群算法优化的 LSSVM 建立预测模型。建模流程如图 4 所示，具体流程描述如下：

(1)根据节目静态特征(演员、情节、导演、片长)，将节目的时间序列数据，分解为 4 种演化类型：外源亚临界类、外源临界类、内源临界类和内源亚临界类。

(2)对 4 种类型的时间序列分别建立 LSSVM 预测模型。采用双种群粒子群算法对影响 LSSVM 预测效果的 3 个参数(惩罚参数 D、核函数参数 σ 以及相关损失函数参数 ε)进行综合选取。

(3)将 4 种模型预测结果按分类概率叠加得到预测流行度。计算 4 种分类在整个节目集中出现的概率 p_k。节目预测流行度 $N_c(t_i, t_r) = \sum_{k=1}^{4} p_k \widehat{N}_{c,k}(t_i, t_r)$。其中，$\widehat{N}_{c,k}(t_i, t_r)$ 为某类节目 LSSVM 模型预测值。

(4)从相对均方根误差和复相关系数 2 个方面对预测结果进行误差分析。

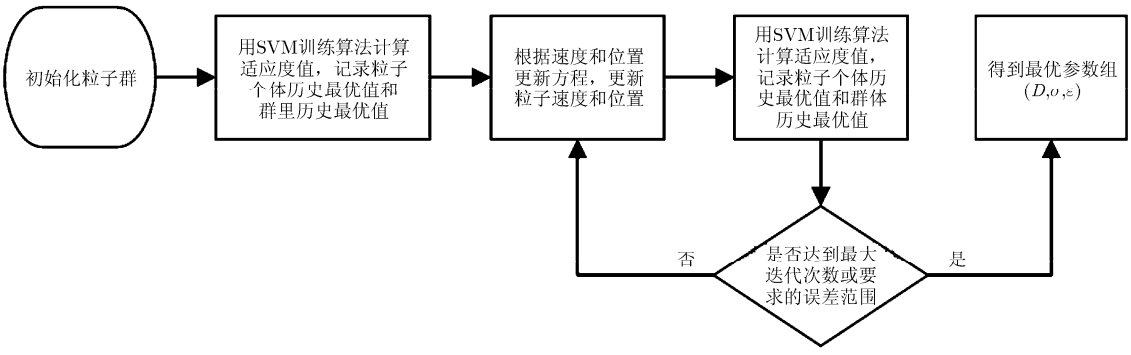


图 3 双种群粒子群优化算法流程图

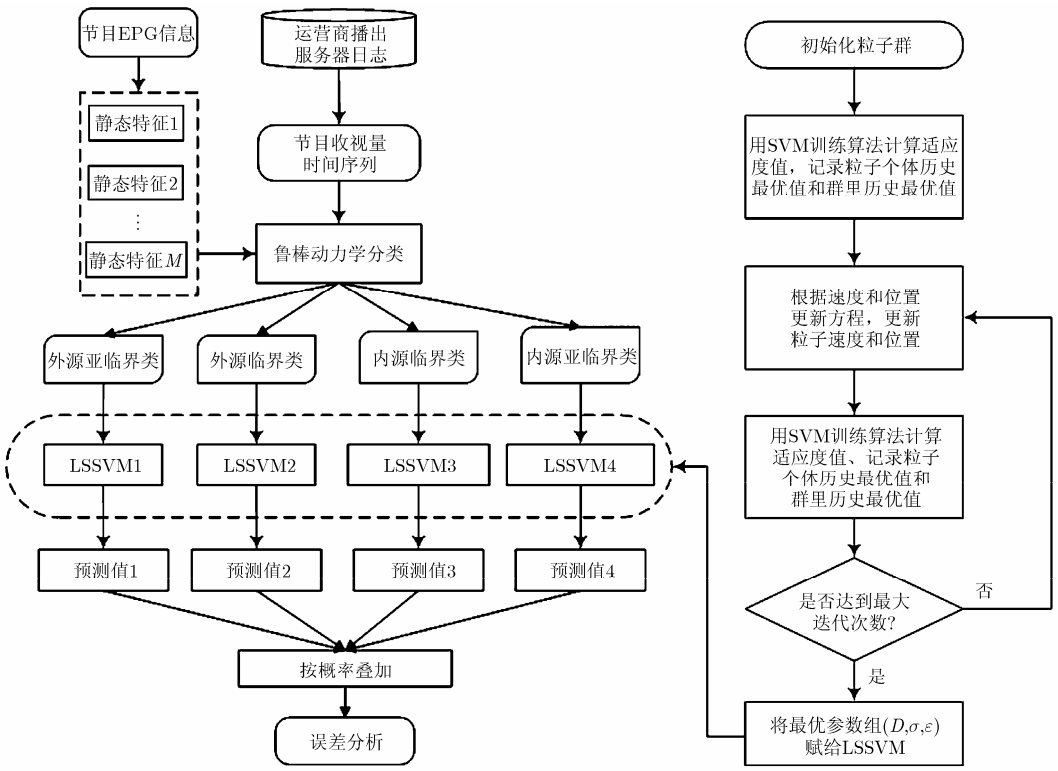


图 4 互联网+电视节目流行度预测流程图

3 实验分析

3.1 数据集介绍

本文的实验数据来源某广电运营商互联网电视平台 2016 年 1 月 1 日至 2016 年 7 月 31 日共计 213 天 280 万用户的 60 亿条收视记录和 120 个频道的电子节目指南数据(Electronic Program Guide, EPG)。通过对视频服务器 RTSP 日志的清洗和对 EPG 信息的解析，分别获得了 11 万部节目的静态特征和流行度时间序列数据。节目静态特征包括导演、编剧、演员、制片国家、语言、播出频道、首播时间、节目时长、节目类型和节目内容概要共 11 项。

3.2 评价指标

全面合理的误差分析可以有效评判预测模型的性能。常用的指标分为绝对误差和相对误差 2 类。采用绝对误差评估模型时需要研究者对预测对象的数值范围有清晰的认识。相对误差采用了误差与实际值的比例，易于实现相同模型在不同数据集上效果的比较，但遇到零值时会造成计算错误。为解决零值问题，可以将预测结果与某个简单模型进行比较，或者直接计算预测值与实际值的复相关系数 R^2 来评估模型的有效性。为实现传统模型与 BD3P 模型在同一数据集上进行比较，规避零值造成的计算错误，本文选取相对均方根误差(Mean Relative Squared Error, MRSE)和复相关系数 R^2 作为评价指标。

RMSE 计算公式如式(16)：

$$\mathrm{RMSE}=\sqrt{\frac{1}{|C|}\sum_{c\in C}\left(\widehat{N}_c\left(t_i,t\right)-N_c\left(t_r\right)\right)^2} \tag{16}$$

R^2 计算公式如式(17)：

$$R^2=1-\frac{\sum_{c\in C}\left(N_c\left(t_r\right)-\widehat{N}_c\left(t_i,t\right)\right)^2}{\sum_{c\in C}\left(N_c\left(t_r\right)-\overline{N}_c\left(t_r\right)\right)^2} \tag{17}$$

式中，C 为预测样本数，$\widehat{N}_c(t_i,t_r)$ 为预测结果，$N_c(t_r)$ 为实测值。

为验证预测模型的输出结果是否一致，本文采用十折交叉验证方法进行了 30 次实验，每次实验将数据集随机平均分成 10 份，轮流选取其中 9 份作为训练数据，1 份作为测试数据，通过检验 30 次实验 RMSE 平均值的标准差 σ 来验证模型预测结果的稳定性。

标准差 σ 计算公式如式(18)：

$$\sigma=\sqrt{\frac{1}{K}\sum_{j=1}^{K}\left(\mathrm{RMSE}_j-\overline{\mathrm{RMSE}}\right)^2} \tag{18}$$

式中，K 为十折交叉验证实验的执行次数。

3.3 结果分析

本文使用 Python 的 Scikit-learn 包实现了 S-H

(Szabo-Huberman)模型和 BD3P 模型，并从相关文献的作者处下载了 ML 和 MRBF 模型的实现代码，将 BD3P 模型与 S-H 模型[6]、ML(Multivariate Linear)模型[7]和 MRBF 模型[8]进行了比较。本文采用前 7 天的流行度数据预测未来 30 天的流行度数据。以 $t_i = 7,\ t_r = 30$ 为例，分别计算了互联网电视平台 112861 个节目采用不同模型预测出的流行度 RMSE。表 1 中记录了 30 次实验的 RMSE 平均值和标准差。

从表 1 数据可以看出，BD3P 模型在整体影片集及所有流行度演化分类中均取得了最佳的 RMSE。在整体影片集上，BD3P 模型相比 S-H 模型预测的 RMSE 约有 30%的降低。特别是对于外源亚临界类的节目，BD3P 模型的 RMSE 约有 80%的降低。这是因为该类节目收视是由现实社会中的突发事件导致(例如热点新闻)，流行度在急剧上升到顶峰后会快速回落。S-H 模型在预测内源临界类节目流行度时取得了较好的精度，但是该模型的线性特性无法准确捕捉外源亚临界类节目流行度衰减的变化趋势。ML 模型根据时间窗口的位置给观测到的流行度数据赋予了不同的权重，一定程度上考虑了演化趋势的方向，降低了预测 RMSE。而 BD3P 模型由于采用 LSSVM，具有更好的逼近能力，因此 RMSE 相比 ML 模型有大约 17.1%的降低。4 种模型的 30 次十折交叉实验 RMSE 的标准差均小于 0.01，预测结果具有较好的稳定性。

MRBF 模型与 BD3P 模型都采用了径向基函数作为回归核函数。径向基函数虽然具有较宽收敛域和泛化能力，但是其性能依赖于超参数的合理选择。MRBF 模型采用网格计算寻找核函数的最优超参数，需要根据经验为参数设定搜索区间。而 BD3P 模型采用双种群的粒子群算法优化超参数，搜索效率更高，同样计算能力下的搜索区间更大。引入两个粒子种群，可以最大程度的避免搜索结果陷入局部最优的情况。因此相比 MRBF 模型，本文提出的算法在核函数寻优性能上有明显优势，反映在 RMSE 上有大约 10%的降低。

本文共收集了 112861 个节目流行度数据，其中属于外源亚临界类的节目有 1554 个，属于外源临界类的节目有 80528 个，属于内源临界类的节目有 6636 个，属于内源亚临界类的有 24143 个。从外源亚临界类、外源临界类和内源临界类中各选取了 1 个节目，基于幂律响应函数参数 θ 的不同取值构建 BD3P 模型，并分析其对预测精度的影响。

《极限挑战》是东方卫视播出的一款综艺真人秀节目，其流行度的变化趋势符合典型内源临界类的特点。虽然节目上线初期收视量较低，但由于节目拍摄常在城市街头，经常与普通观众亲密互动，获得了大量前期收视观众在社交网络的评论和转发，节目流行度逐步提升。在播出 1 周后，流行度达到峰值(下期节目上线前)，并开始逐步递减。图 5 展示了《极限挑战》中不同 θ 取值下，BD3P 预测的内源临界类节目流行度与实际值吻合程度。从图 5 可以看出，θ 取不同值时，BD3P 模型预测精度存在明显的差异；当 $\theta \le 0.3$ 时，预测的峰值流行度远高于实际流行度；当 $\theta \ge 0.45$ 时，流行度初期增长和后期下降速度明显快于实际流行度变化趋势；当 θ =0.4 时流行度峰值和趋势变化速度较符合实际演化趋势，模型取得最佳预测效果。《体育新闻》是一个典型的外源亚临界类的节目。该节目具有较强的时效性，流行度最高峰出现在节目首播时，并且随时间快速下降，上线 7 天后节目的流行度近乎为 0。图 6 展示了《体育新闻》中不同 θ 取值下，BD3P 预测的外源亚临界类节目流行度与实际值吻合程度。

《速度与激情 7》是一个典型的外源临界类节目，虽然流行度最高峰也出现在节目上线初期，但是在上线 20 天后流行度依然达到峰值的 30%。图 7 展示了《速度与激情》中不同 θ 取值下，BD3P 预测的外源临界类节目流行度与实际值贴合程度。从图 6，图 7 可以看出，虽然在预测外援临界类和外源亚临界节目初始流行度时，不同 θ 取值下的模型

表 1 4 种模型 30 次实验的 RMSE 平均值和标准差的比较

	节目数量	S-H[6]		ML[7]		MRBF[8]		BD3P	
		RMSE 均值	标准差	RMSE 均值	标准差	RMSE 均值	标准差	RMSE 均值	标准差
外源亚临界类	1554	0.4743	0.0083	0.1547	0.0082	0.1429	0.0093	0.1177	0.0069
外源临界类	80528	0.2108	0.0013	0.1811	0.0012	0.1731	0.0011	0.1599	0.0011
内源临界类	6636	0.1972	0.0034	0.1863	0.0079	0.1586	0.0054	0.1367	0.0049
内源亚临界类	24143	0.3396	0.0031	0.2989	0.0035	0.2701	0.0035	0.2282	0.0033
整体节目集	112861	0.2403	0.0011	0.2040	0.0013	0.1909	0.0009	0.1692	0.0008

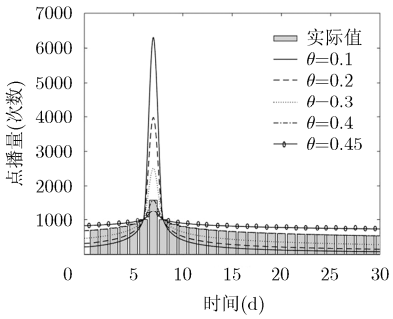


图 5 不同 θ 取值时BD3P模型对内源临界类节目流行度的预测比较

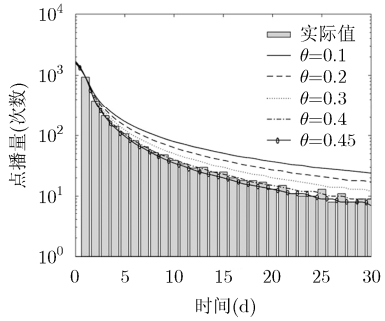


图 6 不同 θ 取值时BD3P模型对外源亚临界类节目流行度的预测比较

的预测精度接近；但随着预测时间的推移，不同 θ 取值模型的预测差异越发明显。与内源临界类节目相同，当 $\theta = 0.4$ 时，模型取得最佳预测效果。

图 8 描述了 4 种模型预测的复相关系数随观测时刻 t_i 影响的情况。随着观测时刻 t_i 变大，可供预测模型训练使用的流行度数据量增多，4 种模型的预测结果与实际值的复相关系数也随之增大。从图 8 可以看出，采用相同的观测数据，BD3P 模型的复相关系数要优于其他 3 种模型。例如，使用 12 天的历史数据($t_i = 12$)，MRBF 模型的复相关系数可以达到 95%。而 BD3P 模型只需要 9 天的数据($t_i = 9$)就能达到相同的效果，缩短了预测周期。

4　结束语

本文从分析传统预测建模存在的不足出发，研究了互联网电视收视行为动力学特征，提出了一种互联网电视节目流行度预测模型 BD3P，并采用双种群粒子群算法提升模型的寻优性能。基于某互联网电视平台的真实用户收视数据，本文将 BD3P 模型与现有模型在预测能力上进行了对比实验，得到如下结论：(1)使用收视行为动力学模型可以有效地描述互联网节目流行度演化过程。根据流行度演化的动力学特征对节目进行分类，减轻了不同趋势信息间的相互影响，有利于深入探究序列特征，从而降低预测误差。(2)与传统的流行度预测模型相比，本文采用的双种群粒子群优化支持向量机具有强劲的全局搜索能力和较快的收敛速度，对提高模型的准确度有较大的贡献。(3)与另外 3 种模型的对比研究验证了 BD3P 模型的有效性。在同等预测精度下，BD3P 模型所需数据量较小，可以缩短预测周期，特别适用于对时间敏感的商业应用环境。

下一步工作拟将 BD3P 模型引入到信息中心网络的缓存调度策略中，尝试建立根据用户收视需求动态适应的缓存替换策略，以提升缓存空间的使用效率，降低网络构建成本。

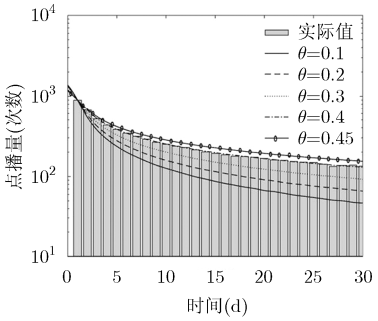


图 7　不同 θ 取值时BD3P模型对外源临界类节目流行度的预测比较

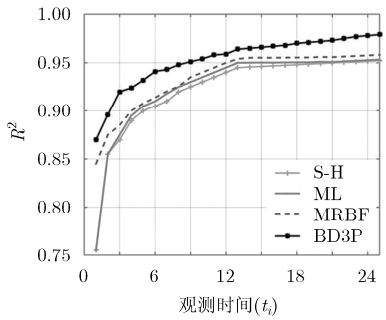


图 8 4种模型预测复相关系数与观测时刻 t_i 的关系

参 考 文 献

[1] 中国互联网络信息中心. 第 38 次中国互联网络发展状况统计报告 [OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201608/P020160803367337470363.pdf.

[2] 腾讯. 腾讯视频电视剧排行榜[OL]. http://v.qq.com/rank/detail/2_-1_-1_-1_1_1.html.

[3] 朱轶, 糜正琨, 王文鼐. 一种基于内容流行度的内容中心网络缓存概率置换策略[J]. 电子与信息学报, 2013, 35(6): 1305–1310. doi: 10.3724/SP.J.1146.2012.01143.

ZHU Yi, MI Zhengkun, and WANG Wennai. A cache probability replacement policy based on content popularity in content centric networks[J]. *Journal of Electronics & Information Technology*, 2013, 35(6): 1305–1310. doi: 10.3724/SP.J.1146.2012.01143.

[4] 芮兰兰, 彭昊, 黄豪球, 等. 基于内容流行度和节点中心度匹配的信息中心网络缓存策略[J]. 电子与信息学报, 2016, 38(2): 325–331. doi: 10.11999/JEIT150626.
RUI Lanlan, PENG Hao, HUANG Haoqiu, *et al*. Popularity and centrality based selective caching scheme for information-centric networks[J]. *Journal of Electronics & Information Technology*, 2016, 38(2): 325–331. doi: 10.11999/JEIT150626.

[5] GÓMEZ V, KALTENBRUNNER A, and LÓPEZ V. Statistical analysis of the social network and discussion threads in slashdot[C]. ACM International Conference on World Wide Web, Beijing, China, 2008: 645–654. doi: 10.1145/1367497.1367585.

[6] SZABO G and HUBERMAN B A. Predicting the popularity of online content[J]. *Communications of the ACM*, 2010, 53(8): 80–88. doi: 10.1145/1787234.1787254.

[7] CASTILLO C, ELHADDAD M, PFEFFER J, *et al*. Characterizing the life cycle of online news stories using social media reactions[C]. ACM International Conference on Computer Supported Cooperative Work & Social Computing, Baltimore, MD, USA, 2014: 211–223. doi: 10.1145/2531602.2531623.

[8] PINTO H, ALMEIDA J M, and GONÇALVES M A. Using early view patterns to predict the popularity of YouTube videos[C]. ACM International Conference on Web Search and Data Mining, Rome, Italy, 2013: 365–374. doi: 10.1145/2433396.2433443.

[9] GAO S, MA J, and CHEN Z. Modeling and predicting retweeting dynamics on microblogging platforms[C]. ACM International Conference on Web Search and Data Mining, Shanghai, China, 2015: 107–116. doi: 10.1145/2684822.2685303.

[10] CRANE R and SORNETTE D. Robust dynamic classes revealed by measuring the response function of a social system[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2008, 105(41): 15649–15653. doi: 10.1073/pnas.0803685105.

[11] WU B, MEI T, CHENG W H, *et al*. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition[C]. Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 2016: 32–38. doi: 10.13140/RG.2.2.27504.66565.

[12] WU J, ZHOU Y, CHIU D M, *et al*. Modeling dynamics of online video popularity[C]. IEEE International Symposium on Quality of Service, Portland, OR, USA, 2015: 141–146. doi: 10.1109/IWQoS.2015.7404724.

[13] FONTANINI G, BERTINI M, and DEL BIMBO A. Web video popularity prediction using sentiment and content visual features[C]. ACM International Conference on Multimedia Retrieval, New York, NY, USA, 2016: 289–292. doi: 10.1145/2911996.2912053.

[14] ZAMAN T, FOX E B, and BRADLOW E T. A Bayesian approach for predicting the popularity of tweets[J]. *The Annals of Applied Statistics*, 2014, 8(3): 1583–1611. doi: 10.1214/14-AOAS741.

[15] WANG J, ZHANG Z, and ZHANG W. Support vector machine based on double-population particle swarm optimization[J]. *Journal of Convergence Information Technology*, 2013, 8(8): 33–43. doi: 10.4156/jcit.vol8.issue8.106.

朱琛刚： 男，1982 年生，博士生，研究方向为 SDN 网络测量、网络流量大数据分析.

程 光： 男，1973 年生，教授，博士生导师，研究方向为 SDN 网络测量、网络流量大数据分析、僵尸网络和 APT 攻击检测.