基于符号回归的产品评论数量与购买数量关系分析

杨 弦[1,2], 党延忠[1], 吴江宁[1]

(1. 大连理工大学系统工程研究所, 辽宁 大连 116024;
2. 东北财经大学管理科学与工程学院, 辽宁 大连 116025)

摘要: 为揭示产品评论数量与购买数量间的关系, 采用数据驱动的符号回归方法, 该方法不需预先假设模型结构, 从数据中自动学习出描述产品评论数量与购买数量的关系模型和相关参数. 数据采自天猫商城网站中隶属于 30 种商品的 5 387 个售出产品以及相关评论, 利用符号回归方法进行自动学习, 发现了在实证方法中假设的线性模型. 这种模式在购买数量与评论数量之间的关系上具有更高的适应性.

关键词: 购买数量; 评论数量; 关系模型; 符号回归方法
中图分类号: C931 **文献标识码**: A **文章编号**: 1000−5781(2020)03−0289−12
doi: 10.13383/j.cnki.jse.2020.03.001

Analysis on the relationship between the number of product reviews and the number of purchases based on symbolic regression

Yang Xian[1,2], Dang Yanzhong[1], Wu Jiangning[1]

(1. Institute of Systems Engineering, Dalian University of Technology, Dalian 116024, China;
2. School of Management and Science Engineering, Dongbei University Finance and Economics, Dalian 116025, China)

Abstract: In order to reveal the relationship between the volume of purchases and the volume of reviews, this paper introduces a data-driven symbolic regression method to discover both the form of the model and its parameters intelligently based on a large amount of data. Compared with previous empirical studies, symbolic regression via genetic programming does not need a prior hypothetical model form to fit the observed data. The data sets of this paper are collected from Tmall.com. The data sets include 5 387 products belonging to 30 categories. By the method, the linear model assumed in the prior empirical methods is found. Meanwhile, new nonlinear models are also found. Such models hasve a better fitness in terms of the relationship between the volume of purchases and the volume of reviews.

Key words: volume of purchases; volume of consumer reviews; relationship model; symbolic regression method

1 引 言

为获得消费者的反馈信息, 电子商务网站如亚马逊、京东、天猫商城和当当网等大都自设评论平台, 鼓励消费者购买产品后在此发表评论, 分享自己的购物体验. 当前, 在线评论已成为电子商务网站最受欢迎

收稿日期: 2018−04−13; 修订日期: 2018−10−22.
基金项目: 国家自然科学基金资助项目(71871041; 71671024); 大连市科技创新基金资助项目(2108J11CY009); 中央高校基本科研业务费资助项目(DUT20JC38; DUT20WR301).

的、不可或缺的属性之一. 在线评论作为一种用户生成内容(user generated content, UGC), 是用户反馈的产品口碑信息, 它的数量反映了产品在市场中的热度, 不仅对产品销量有影响[1,2], 还会影响营销策略[3−7]. 在线评论既是消费者购买产品的驱动力, 也是消费者购买并体验产品后的产出. 购买者越多, 产出的评论也越多[8,9]. 尽管评论数量与购买数量呈正相关关系, 但二者间的函数关系却不得而知.

关于二者函数关系的研究有着很重要的意义, 这是由于出于对商业秘密的保护, 大多数电子商务网站都选择不公开其产品的销量数据. 因此在研究评论内容以及其他因素对产品在线销量的影响时, 只能使用公开的产品评论数量数据代替隐藏的历史销量数据做进一步研究[10−13]. 这些研究大都根据主观判断, 事先假设评论数量与购买数量间是简单的线性关系, 即评论数量 $= a\times$ 购买数量. 如 Ye 等[10,11]在研究携程网在线评论的各属性对酒店客房销量的影响时, 他们假定评论数量与历史客房销量之间存在线性关系, 且相关系数 a 为常数, 并用评论数量代替历史客房销量做进一步研究. Lee 等[12] 在研究亚马逊网站上产品购买数量的分布时, 也假定评论数量与购买数量呈一定比例的线性关系, 并用评论数量代替购买数量. 王君珺等[13]用京东商城中手机的在线评论数量代替产品购买数量, 进一步研究相关因素对产品在线购买数量的影响. 这种主观判断和假设没有被相关研究以完备的数据明确验证过, 如果假设的线性关系不成立, 相关研究结果将会出现测量偏差.

那么, 评论数量与购买数量间到底为何种关系, 其结构模型是什么, 这些问题仍值得深入研究. 目前的研究存在以下两方面不足: 1) 数据的不完备性. 之前的相关研究都只有评论数量的数据, 没有对应的产品销量数据, 这使得学者们只能假设替代关系, 用公开的评论数据代替隐藏的销售数据, 常用的方法是事先假定线性关系模型. 2)预设模型的局限性. 通过定性分析假定关系模型, 往往受领域知识和认知水平的限制, 所构建的模型不能完全刻画评论数量与购买数量间的复杂关系.

因此, 本文在同时获取了产品的评论数量和购买数量的完备数据的基础上, 研究评论数量与购买数量间精确的定量关系, 可为之前用评论数量代替购买数量的研究提供科学依据. 此外, 在关系模型的构建过程中, 不预设模型, 采用了一种智能的模型构建方法, 一方面验证了以往研究中依据主观判断所假设的模型, 另一方面还发现了其他多个新的模型, 这相当于构建了一个模型库. 在实际应用中, 可根据不同指标如精度、复杂度和适应度等, 以及模型各自的优缺点, 对模型进行综合判断和选择.

本文的研究数据来自天猫商城, 该网站同时提供了海量的评论数据和购买数据信息, 为研究评论数量与购买数量间的相关关系提供了完备的数据支持. 针对以往研究的不足, 具体提出以下两个研究问题: 产品评论数量与购买数量之间是否是线性关系? 是否存在更为复杂的非线性关系? 此研究旨在通过多种关系模型的构建, 发现两者之间的各种关系. 为发现评论数量与购买数量的关系模型, 直接从数据出发, 采用一种新颖的基于数据驱动的符号回归方法自动学习关系模型和参数. 实验结果不仅验证了已有的假设, 还发现了评论数量与购买数量间更为复杂的关系.

本文在理论方面, 提供了一种关系模型构建和模型选择的新思路和方法. 该方法有效弥补了提前假设模型方法的不足, 结果更客观和丰富. 在应用方面, 研究结果可启发商家根据产品的不同购买数量制定相应的奖励政策来刺激消费者分享购物体验, 以增强产品的口碑效应, 实现销量与口碑齐升的双重目标.

2 研究思路和方法

2.1 研究思路

本文不基于关系假设进行模型构建, 直接从数据出发, 采用基于数据驱动的符号回归方法自动地学习模型以及参数. 符号回归方法[14]不需要假定任何函数形式, 以达尔文自然选择进化理论为依托, 利用计算机程序模拟基因复制、交叉和变异等操作, 优胜劣汰, 从数据中自动地发现知识、模式和规律.

按照数据获取→变量定义→模型构建→关系发现的逻辑顺序开展工作, 如图 1 所示. 首先, 通过爬虫软件在天猫商城网站中收集产品购买数量和相关评论数据; 然后, 对数据进行预处理, 选择产品评论数量和购买数量两个变量; 接着, 通过符号回归方法发现适应于选定产品的关系模型, 变量之间的具体函数形式; 最

后, 基于模型的函数形式分析变量之间的关系.

在研究评论数量与购买数量之间的关系时, 只选取了两个变量, 虽然过于简单, 但其意义和合理性是明显的. 旨在探究评论数量与购买数量二者间的精确定量关系, 并验证之前相关研究的研究假设(评论数量 $=a\times$ 购买数量) 是否存在, 以及是否存在更复杂的关系. 因此, 研究中只考虑了这两个变量, 识别变量间简单的基本关系, 并且用具体的函数模型来表示这种关系, 在市场营销科学的相关研究中是很有意义的[15]. 这些基础模型更具普适性, 且能为将来的相关研究重复和扩展变量之间的关系提供研究起点[16].

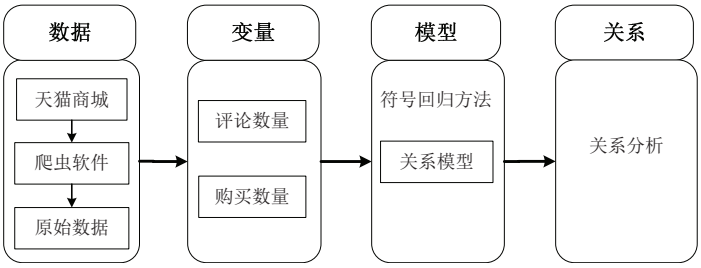


图 1　研究思路

Fig. 1　Research framework

2.2　符号回归方法

符号回归方法最早由 Koza[14]提出, 其主要思想是依据达尔文的进化论, 自动更新模型结构及参数以获得最优化模型. 传统的回归方法[17], 首先需要基于领域知识定性分析变量之间的可能关系, 定义函数形式, 如线性或非线性、一次或多项式等, 然后基于观测数据, 估计所定义函数实行的最优参数. 与之不同的是, 符号回归方法不需模型结构假设, 能够从数据中学习出模型结构及其参数[18]. 也就是说, 当研究者对数据生成系统的领域知识有限、较难确定模型结构时, 基于数据驱动的符号回归是一种行之有效的方法. 符号回归作为一种函数发现方法, 能够自动描述变量之间的关系, 并以数学函数的形式予以表示[19].

符号回归方法现已被用于很多领域来探索变量之间的内在关系. 基于符号回归, Schmidt 等[20]在不预先导入任何物理学、几何学和动力学等领域知识的前提下, 通过分析记录的运动轨迹数据, 用计算机程序自动发现了汉密尔顿函数、拉格朗日函数和动量守恒定律等. Chattopadhyay 等[21]将该方法应用于生物领域, 他们从一系列的系统观测数据中, 自动学习出成百上千的反应模型. Kemp 等[22]针对传统的数据建模存在只能预先固定模型结构的局限性, 提出一套新思路, 从数据中同时学习多个候选模型. Yang 等[23]通过符号回归方法, 探索环境污染与经济发展的关系, 他们通过解析世界各国 CO_2 和 GDP 的数据, 智能地学习出包括 EKC 模型在内的所有六种经典规律模型.

利用符号回归方法探索产品评论数量与购买数量的关系, 在不事先假设模型结构的前提下, 从大量的观测数据中找出关系模型的结构及其参数. 符号回归过程主要执行的是遗传算法, 其步骤如下:

步骤 1　随机产生初始种群;

步骤 2　设置终止条件, 迭代执行步骤 3 和 步骤 4 直到满足终止条件;

步骤 3　利用适应度函数计算个体的适应度值;

步骤 4　按照一定的复制概率保留个体到下一代, 按照一定的交叉概率和随机交叉操作形成下一代新的个体, 按照一定的变异概率和随机变异操作形成下一代新的个体;

步骤 5　选择适应度最佳的个体.

在符号回归的进化过程中, 候选解决方案一般以树形结构表示. 例如, $0.5x+0.2x^2+3$ 可由图 2 所示的树形结构表示. 通常越复杂的候选模型其精度越高, 但可能出现“过度拟合”现象. 为了衡量候选模型的复杂度, 文中引进了复杂度指标 C, 定义四则运算符的复杂度为 1, 常数和变量的复杂度为 1, exp 和 ln 的复杂度为 4. 图 2 所示的例子中, 表达式的复杂度为 11.

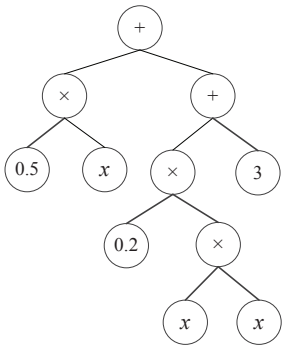


图 2 表达式 $0.5x+0.2x^2+3$ 的树形结构图

Fig. 2 Tree structure for the expression $0.5x+0.2x^2+3$

算法中个体的进化主要执行三种遗传操作: 复制、交叉和变异. 复制操作: 选择最好的个体直接到下一代; 交叉操作: 交叉两个个体的部分特征得到两个新的个体; 变异操作: 改变个体的一部分, 得到一个新的个体. 个体进化的遗传操作均在树形结构上进行(如图 3 所示).

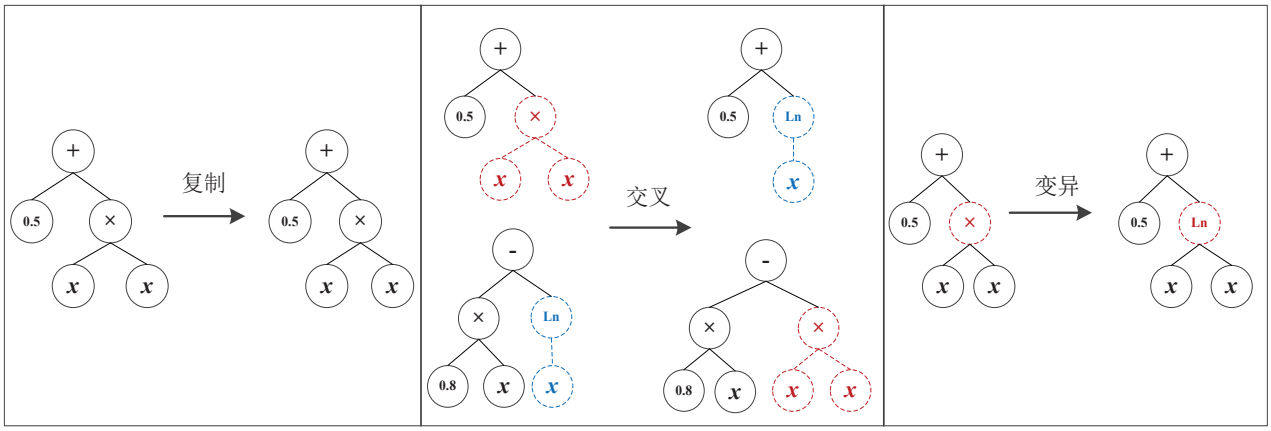


图 3 遗传操作示意图

Fig. 3 The genetic operation schematic diagram

利用符号回归方法生成模型的过程如下:

步骤 1 选择一组算术、函数运算符和一组变量. 选择的算术运算符包括加(+)、减(−)、乘(×)和除(÷); 函数运算符包括指数函数(exp)和对数函数(ln); 变量包括产品的评论数量和购买数量. 设置程序的运行时间为 1 000 s, 此时, 输出的模型拟合效果稳定可靠[1]. 换言之, 即便增加再多的运算时间, 拟合效果也没有很大的改善空间;

步骤 2 进行遗传操作, 通过复制、交叉和变异来进化模型的结构和参数. 每一次进化后生成模型的同时, 输出该模型的复杂度 C 和模型的拟合优度 R^2; 拟合优度 R^2 用来测量个体的质量, 其计算公式为

$$R^2=1-\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2\left(\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2\right)^{-1}, \tag{1}$$

其中 Y_i 表示因变量的实际值, $\hat{Y}_i$ 表示因变量的预测值, $\bar{Y}_i$ 表示因变量的平均值, n 是数据的总数;

步骤 3 评价模型. 经过遗传操作, 可输出大量的候选模型. 用拟合优度和复杂度来评价所获得的模型, 在确保模型质量的前提下选择最优模型组. 具体的, 用 R^2 阈值过滤掉拟合优度较低的模型, 通过最大复杂度 $C_{\max}$ 预防过拟合. 依据奥卡姆剃刀定律(Occam's Razor Law), 如果两个模型有同样的精度, 则选择复

[1]遗传算法在进行模型迭代时, 迭代耗费时间在 1 000 s以下进化结果已经收敛, 实验中发现, 大部分过程达到收敛态耗时低于 500 s. 通常, 算法运行时长与问题的复杂程度, 数据规模有关, 问题越复杂, 数据量越大, 算法的收敛时间越长[20]. 考察的问题复杂程度和数据量均不是很大, 符号回归可以较快地达到收敛状态.

杂度低的模型, 选择简单而有效的模型作为输出. 此外, 用帕累托前沿(Pareto front) 平衡模型的精度和复杂度[20,23]基于帕累托前沿, 当模型有同样的复杂度时, 程序将选择精度较高的模型; 当模型有同等的拟合精度时, 程序将选择复杂度较低的模型. 最终, 构建了一个帕累托前沿, 从众多的候选模型中选出了满足要求的模型组.

因此, 针对一组给定数据集, 符号回归方法可以输出多个模型来解释评论数量与购买数量之间的关系. 这是由于对于某一数据集, 其内在特征可从多方面, 多角度观察, 角度不同对应的模型亦不同. 传统的统计分析方法, 常用一种模型刻画数据特征, 而符号回归方法可同时发现多个规律模型, 其中某一个模型可理解为数据集的一个剖面所表现出的特征[23]. 这也是符号回归方法的优点之一, 该方法不仅能输出一个具有较高精度的模型, 且能发现多个各有优劣的模型, 不同模型从不同视角来看都具有一定的合理性[23].

3　评论数量与购买数量关系分析实验

3.1　数据采集

实验所用数据包括产品的相关评论的数量和购买数量全部来自天猫商城网站(www.tmall.com). 天猫商城是阿里巴巴旗下中国最大的 B2C 网站, 有别于其他的 B2C 网站, 如亚马逊、京东和当当等网站, 它们一般不提供产品的销量信息, 天猫商城店铺的主页面同时提供了产品的评论数量和购买数量信息, 这为本文的研究提供了大量可信且有效的数据支持. 实验选取了 30 个产品种类, 利用开发的爬虫软件收集了这些产品种类下的 5 387 个商品的评论数量和购买数量数据, 其描述性统计结果见表 1.

表 1　数据的描述性统计结果

Table 1　Descriptive statistical results of the data

产品种类	产品数量 / 个	购买数量 / 个				评论数量 / 条			
		最小值	最大值	平均值	标准差	最小值	最大值	平均值	标准差
家电与电子产品类									
手机	104	8	208 388	15 510.2	33 959.3	0	92 081	5 252.3	12 635.4
手提电脑	200	1	18 452	725.6	2 079.9	0	9 807	320.4	992.6
耳机	191	1	72 751	3 171.1	8 406.6	0	30 797	1 200.5	3 219.5
U盘	186	1	944 527	16 707.1	85 740.5	0	254 354	4 981.7	23 526.7
路由器	158	2	2 562 409	55 513.1	232 561.8	0	871 523	17 588.2	78 275.2
照相机	158	0	5 038	250.8	583.8	0	1 639	84.1	194.2
洗衣机	172	1	148 688	8 110.2	18 481.7	0	68 498	3 440.9	7 856.7
冰箱	191	1	210 819	7 701.5	23 615.0	0	64 582	2 835.0	7 181.5
电视机	200	1	69 207	3 928.1	8 125.4	0	26 811	1 779.4	3 550.6
空调	188	1	90 963	6 163.4	12 037.8	0	35 496	2 384.2	4 730.8
纺织品与服装类									
T恤	200	3	47 210	919.4	3 561.6	0	15 901	257.0	1 192.7
连衣裙	200	20	8 177	915.5	1 343.5	0	2 199	235.7	371.1
牛仔裤	200	1	25 637	742.8	2 523.3	0	7 873	233.0	822.2
西服	193	0	46 962	761.3	4 224.2	0	9 757	210.5	881.4
运动鞋	166	1	16 159	1 031.5	2 345.9	0	5 702	360.9	862.5
单鞋	199	1	18 625	1 620.4	3 141.5	0	6 715	588.6	1 172.4
行李箱	200	0	72 072	2 288.1	6 424.6	0	11 155	651.2	1 399.4
单肩包	200	0	7 001	648.2	1 116.5	0	2 967	226.1	418.3
四件套	180	1	40 358	2 027.0	5 821.6	0	14 652	424.4	1 346.0
枕头	80	8	117 909	5 590.8	16 952.0	1	38 072	1 334.0	4 788.2
食品与生活用品类									
白酒	199	4	212 201	5 495.6	17 531.4	0	45 152	1 205.7	3 807.4
茶	200	0	408 700	12 337.5	41 644.5	0	101 105	2 513.8	9 357.1
巧克力	200	0	683 057	16 128.1	55 579.7	0	157 354	3 160.1	12 082.4
坚果	200	32	2 665 548	566 061.2	602 581.9	9	728 840	161 855.0	169 194.4

表 1 (续)
Table 1 (Continue)

产品种类	产品数量 / 个	购买数量 / 个				评论数量 / 条			
		最小值	最大值	平均值	标准差	最小值	最大值	平均值	标准差
食品与生活用品类									
大米	114	0	85 799	3 179.4	10 333.0	0	14 242	524.7	1 809.5
沐浴露	200	1	313 912	12 516.9	32 685.3	0	100 239	4 468.2	11 354.4
牙膏	195	0	634 695	19 213.5	67 226.6	0	148 885	5 397.1	17 633.6
香水	200	0	54 433	1 271.2	4 592.2	0	19 877	461.1	1 671.2
口红	108	0	186 050	5 058.4	23 868.3	0	65 674	2 064.3	9 641.2
太阳镜	205	0	42 645	1 603.8	5 764.8	0	10 843	529.6	1 749.1
汇总	5 387			25 906.4				7 552.3	

由表 1 可知, 各产品种类的评论数量总体上小于购买数量, 且平均评论数量(7 552.3)是平均购买数量(25 906.4)的三分之一左右.

3.2 实验结果分析

3.2.1 通用的关系模型

实验中, 每一个产品种类的数据形成一个数据集, 针对给定数据集利用符号回归方法可得到一组帕累托最优模型组. 因此, 对于30个产品种类数据得到了 30 个帕累托最优模型组. 为考察是否存在通用的关系模型适合每一个产品种类, 实验中合并统计了所有产品种类的帕累托最优模型组, 依据 r 指标值进行降序排列, 排名前 10 的模型列于表 2.

表 2 前 10 个按解释度降序排列的关系模型
Table 2 Top 10 relationship models based on the descending r

模型	函数形式	C	$\overline{R^2}(D)$	r
M1	$v_{\mathrm{r}} = av_{\mathrm{p}}$	3	0.846(0.203)	100.00%
M2	$v_{\mathrm{r}} = av_{\mathrm{p}} - bv_{\mathrm{p}}^2$	9	0.900(0.102)	60.00%
M3	$v_{\mathrm{r}} = av_{\mathrm{p}} - b$	5	0.911(0.172)	53.33%
M4	$v_{\mathrm{r}} = av_{\mathrm{p}} - b - cv_{\mathrm{p}}^2$	11	0.927(0.080)	26.67%
M5	$v_{\mathrm{r}} = av_{\mathrm{p}} + bv_{\mathrm{p}}^3 - cv_{\mathrm{p}}^2$	17	0.811(0.190)	23.33%
M6	$v_{\mathrm{r}} = a + bv_{\mathrm{p}}$	5	0.685(0.152)	23.33%
M7	$v_{\mathrm{r}} = av_{\mathrm{p}} + bv_{\mathrm{p}}^2 - cv_{\mathrm{p}}^3$	17	0.949(0.073)	20.00%
M8	$v_{\mathrm{r}} = av_{\mathrm{p}} + bv_{\mathrm{p}}^2 + cv_{\mathrm{p}}^4 - dv_{\mathrm{p}}^3$	27	0.964(0.054)	16.67%
M9	$v_{\mathrm{r}} = av_{\mathrm{p}} - bv_{\mathrm{p}}^3$	11	0.930(0.078)	16.67%
M10	$v_{\mathrm{r}} = av_{\mathrm{p}} - \ln(v_{\mathrm{p}})$	9	0.961(0.064)	13.33%

这里, r 定义为模型的解释度, 是指模型能够解释产品种类的比例, 值越大表示模型的适用范围越广, 其计算公式为

$$r_j = \frac{n_j}{m}, \tag{2}$$

其中 m 表示产品种类的总数量, n_j 表示所发现模型 j 的产品种类数量.

在表 2 中, $\overline{R^2}$是拟合优度 R^2 的平均值; D 是 R^2 的标准差; C 为模型的复杂度; v_{p} 表示产品的购买数量(volume of purchase); v_{r}表示评论数量(volume of review); 模型中所有变量前的系数(如 a, b, c, d 等)均为正数. 表中排名越靠前的关系模型可解释的产品种类越多, 模型的适用范围越广. 排名前三的模型 M1, M2 和 M3 可解释 50 % 以上的产品种类, 解释能力较强, 且具有较高的平均拟合优度($\overline{R^2}$值都大于 0.8).

为考察候选模型的鲁棒性, 观察不同模型在不同 R^2 阈值的条件下是否稳定, 实验中分别设置不同的 R^2 阈值给出模型分布. R^2 阈值是筛选模型时设定的 R^2 需满足的最小值, 用于保留高于 R^2阈值的模型, 去掉低于 R^2 阈值的模型. 在每一个 R^2 阈值设定条件下, 筛选出符合条件的模型, 然后利用 r 对模型降序排列, 排序结果见表 3.

表 3　不同 R^2 阈值对应的模型分布

Table 3　The distribution of models corresponding to different thresholds of R^2

r 降序	$R^2>0.1$	$R^2>0.2$	$R^2>0.3$	$R^2>0.4$	$R^2>0.5$	$R^2>0.6$	$R^2>0.7$	$R^2>0.8$	$R^2>0.9$
1	M1	M1	M1	M1	M1	M1	M1	M1	M1
2	M2	M2	M2	M2	M2	M2	M2	M3	M3
3	M3	M3	M3	M3	M3	M3	M3	M2	M2
4	M4	M4	M4	M4	M4	M4	M4	M4	M4
5	M5	M5	M5	M5	M5	M5	M7	M7	M7
6	M6	M6	M6	M6	M7	M7	M8	M8	M8
7	M7	M7	M7	M7	M6	M6	M5	M9	M10
8	M8	M8	M8	M8	M8	M8	M9	M10	M6
9	M9	M9	M9	M9	M9	M9	M10	M5	M9
10	M10	M10	M10	M10	M10	M10	M6	M6	M5

例如, 当 R^2 阈值是 0.5 时, 模型 M1∼M10 的所有 R^2 值都大于 0.5, 这 10 个模型的解释度 r 由大到小依次为 M1, M2, M3, M4, M5, M7, M6, M8, M9 和 M10. 从表 3 中可以看出, 当 R^2 阈值在不同的取值范围时, 三个模型 M1, M2 和 M3 的排序虽略有变化, 但都比较稳定地出现在前三的位置上, 说明模型 M1, M2 和 M3 具有较强的鲁棒性.

接下来, 主要针对模型 M1, M2 和 M3 的所有不同 R^2 值的函数进行分析. 为了可视化展现这三个模型对原始数据的拟合情况, 实验中分别选取模型 M1, M2 和 M3 中对应的前三个 R^2 值最高的函数, 拟合曲线如图 4 所示.

由三个模型的函数形式可知, 模型 M1 和模型 M3 是线性模型, 模型 M2 是非线性模型, 下面分别详细分析三个模型的函数形式及其所表达的评论数量与购买数量之间的关系.

1) 线性模型

模型 M1 能够以 $\overline{R^2}$ 值为 0.846 的平均拟合优度解释所有产品种类的评论数量与购买数量间的关系. 该模型的函数形式为 $v_{\mathrm{r}}=av_{\mathrm{p}}$, 其中 a 是斜率, $a>0$ 表示评论数量与购买数量正线性相关, 亦即评论数量以常数 a 的速率随着购买数量的增加而增加. 即产品的评论密度[2] 是稳定不变的, 与购买数量无关. M1 模型与之前相关研究[10−13]中的线性模型基本一致, 本文用真实数据验证了线性假设的存在, 这一研究结果为评论数量与购买数量间存在替代关系提供了科学依据. 模型 M1 是最简单的线性模型(复杂度为 5), 适用范围广, 但它的平均拟合优度 $\overline{R^2}$ 在三个模型中略低.

模型 M3 的函数形式为 $v_{\mathrm{r}}=av_{\mathrm{p}}-b$. 与模型 M1 不同的是模型 M3 带有截距. 可以理解为当产品的购买数量大于 a/b 时, 产品的评论数量将大于 0, 如图 4(c)所示. 模型 M3 能够以 $\overline{R^2}$ 值为 0.911 的平均拟合优度解释 53.3 % 的产品种类, 即 16 个产品种类中评论数量与购买数量间的关系可由模型 M3 刻画. 利用模型 M3 的 16 个具体函数的系数, 计算 b/a 的平均值取整后为 14. 由此可知, 53.3 % 的产品种类中, 只有当产品的购买数量平均超过 14 个时才会有评论. 因此, 对于新进入产品, 商家可以考虑对前 14 个购买该产品的消费者给予更多的优惠, 以刺激他们撰写评论[25].

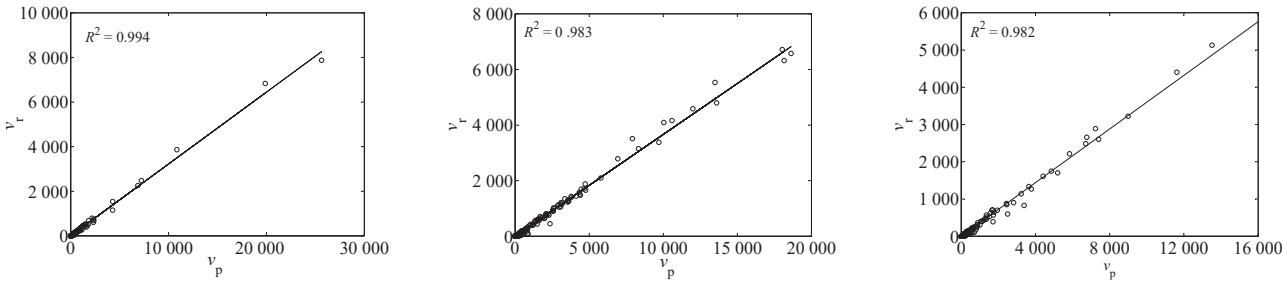
2) 非线性模型

除了线性模型, 通过符号回归方法还发现了新的非线性模型 M2. 该模型能以较高的平均拟合优度 $\overline{R^2}$ 值 0.900 解释 60 % 的产品种类, 即在 18 个产品种类中的数据发现了该模型. M2 模型的函数形式为 $v_{\mathrm{r}}=av_{\mathrm{p}}-bv_{\mathrm{p}}^2$, 它是一个顶点坐标$(v_{\mathrm{p0}},v_{\mathrm{r0}})$为$(a/(2b),a^2/(4b))$的二次函数.

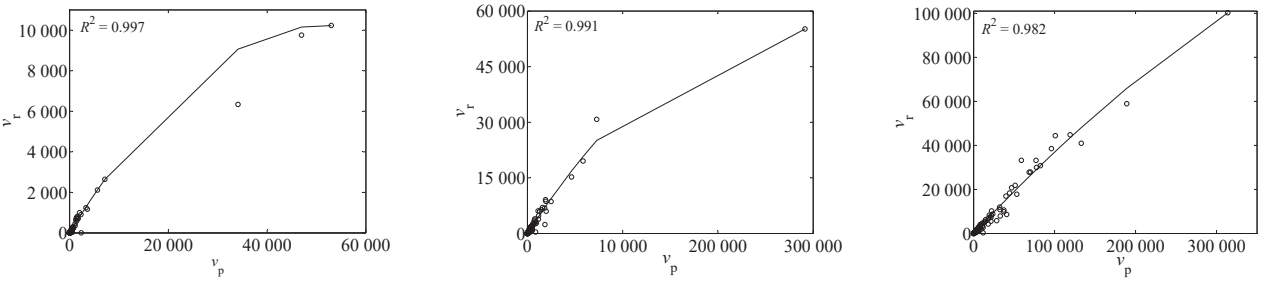
为了分析该模型所刻画的具体关系, 基于符合该模型形式的 18 个函数的系数, 计算各函数的顶点横坐标值 $v_{\mathrm{p0}}^i,i=1,2,\ldots,18$, 同时统计 18 个产品种类数据中横坐标的最大值 $v_{\mathrm{p\,max}}^i,i=1,2,\ldots,18$, 比较两者大小可知, $v_{\mathrm{p\,max}}^i<v_{\mathrm{p0}}^i,i=1,2,\ldots,18$, 见图 5.

[2]评论密度是指消费者群体购买产品后是否发表评论的总体倾向. 相关研究[24] 定义评论密度是在一个给定的时间内, 对于某产品发表评论的数量与产品购买数量的比值.

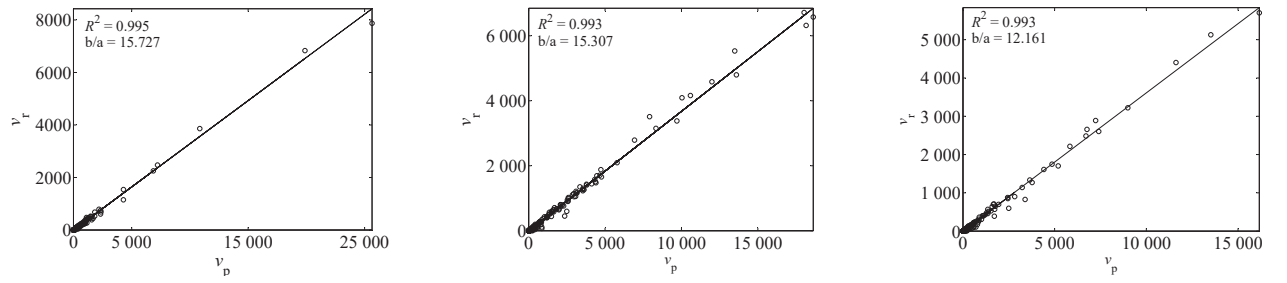
从模型 M2 表示的关系可知, 尽管 M2 是二次函数, 但只有一部分曲线(图 5 中的实线)拟合了原始数据, 可以看出模型 M2 在 $v_{\mathrm{p}} \in [v_{\mathrm{p}\min}, v_{\mathrm{p}\max}]$上是单调函数, 说明评论数量 v_{r} 随着购买数量 v_{p} 的增加而增加, 但增加的速率在降低. 该现象表示购买数量较少的产品, 评论密度相对较高; 而购买数量较大的产品, 评论密度会相对较低. 关系模型 M2 在之前的研究中没有被提出过, 它所表示的评论数量与购买数量的关系可以用消费者行为理论给予解释.



(a) M1: $v_{\mathrm{r}} = av_{\mathrm{p}}$



(b) M2: $v_{\mathrm{r}} = av_{\mathrm{p}} - bv_{\mathrm{p}}^2$



(c) M3: $v_{\mathrm{r}} = av_{\mathrm{p}} - b$

图 4 模型 M1, M2 和 M3 的前三个 R^2 最高的函数拟合曲线

Fig. 4 The top three highest R^2 fitting curves for M1, M2, and M3

消费者行为的相关文献[26,27]指出, 由于消费者传播口碑是利他和互惠的, 为了自我增强, 以及表达独特性, 那些销量低的产品反而会吸引更多的消费者去评论. Shen 等[28]对亚马逊网站中图书的研究发现, 那些畅销且获得较多评论的图书, 消费者总体的评论倾向会减低. 因此引发了产品购买数量低评论密度高, 或者购买数量高评论密度低的这种非常规现象[3].

[3]文献[16–28]的结论能解释 M2, 但与模型 M1 的结论相悖. 主要是由于文献[16–18]的结论体现的是一种规律, 这在本文的模型结果中得到体现, 但这种规律有其适用范围, 并不是所有产品数据都会严格遵循这种规律. 换言之, 在其适用范围内, 这种规律模型的表现是优异的, 但超出其适用范围, 其表现将会出现较大偏差. 模型 M1 的解释数据的范围相对要广一些, 超出了文献[26–28]的解释范围, 因此出现了相悖结论是可以理解的.

本文发现的非线性模型 M2 能以更高的平均拟合优度解释评论数量与购买数量之间复杂的关系, 更精确地刻画了消费者购买产品后的总体评论倾向. 该模型带给商家的启示是, 对于购买数量大的产品应考虑提供更为优厚的评论奖励, 提高评论密度, 最大化产品评论的总数量.

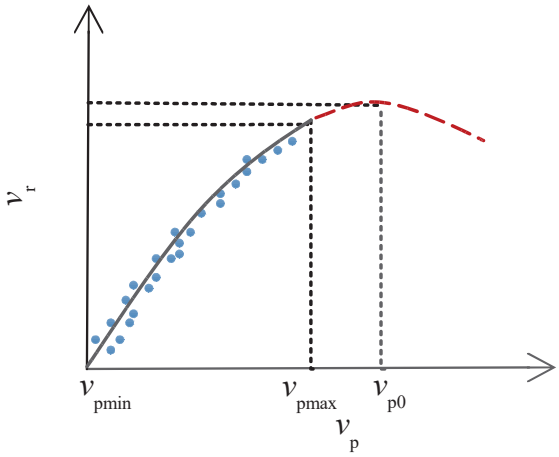


图 5　模型 M2 表示的关系

Fig. 5　The relationship expression of M2

3.2.2 不同产品类型的关系模型

本文对所有产品种类进行了细分, 比较分析不同产品类型对应的关系模型的适用范围. 根据本文收集的30个产品种类, 将所获产品数据分成了家电与电子产品、纺织品与服装和食品与生活用品三大产品类型[4], 每种类型中分别包含 10 个产品种类, 具体的分类如表 1 所示. 不同产品类型下的所有关系模型及其相应的指标值列于表 4, 表中模型是按 r 值降序排列的, 排序越靠前表示模型能解释该产品类型下的产品种类数量越多, 模型对该产品类型的适用范围越广[5]. 例如, 在家电与电子产品类中, 模型 M1 以 0.845 的平均拟合优度(标准差为 0.145), 来解释 100 % 的产品种类, 该模型的适用范围最广.

表 4　不同产品类型的关系模型

Table 4　Relationship models of different product types

r 降序	家电与电子产品类			纺织品与服装类			食品与生活用品类		
	模型	$\overline{R^2}(D)$	r	模型	$\overline{R^2}(D)$	r	模型	$\overline{R^2}(D)$	r
1	M1	0.845(0.145)	100.00%	M1	0.854(0.208)	100.00%	M1	0.841(0.261)	100.00%
2	M2	0.928(0.073)	70.00%	M3	0.975(0.031)	60.00%	M2	0.935(0.062)	70.00%
3	M3	0.884(0.047)	40.00%	M2	0.791(0.008)	40.00%	M3	0.865(0.267)	60.00%
4	M4	0.938(0.116)	40.00%	M6	0.698(0.142)	40.00%	M9	0.930(0.078)	50.00%
5	M5	0.889(0.129)	20.00%	M7	0.967(0.051)	30.00%	M5	0.808(0.273)	30.00%
6	M8	0.926(0.080)	20.00%	M10	0.993(0.001)	30.00%	M8	0.989(0.009)	30.00%
7	M6	0.623(0.007)	20.00%	M4	0.874(0.175)	20.00%	M7	0.932(0.098)	30.00%
8	M10	0.865(0.000)	10.00%	M5	0.738(0.166)	20.00%	M4	0.957(0.006)	20.00%

通过对不同产品类型下各模型的适用能力进行比较分析可知, 简单的线性模型 M1 在三个产品类型中都能以高于 0.840 的平均拟合优度解释所有的产品. 因此, 模型 M1 对产品类型的适用范围没有明显的不同. 而带截距的线性模型 M3 相比于在其他产品类型中的适用能力, 在纺织品与服装类中其平均拟合优度更高, 适用范围更广. 非线性模型 M2 对家电与电子产品类和食品与生活用品类都能以高于 0.920 的平均拟合优

[4]分类标准主要是基于国家统计局的统计用产品分类目录.

[5]表 4 与表 2 比少了两个模型, 是由于特定产品类型下的产品没有发现 M1∼M10 中的某两个模型.

度解释 70 % 的产品, 而对于纺织品与服装类只能以平均 0.791 的平均拟合优度解释该产品类型下的 40 % 产品. 因此, 非线性模型 M2 更适用于家电与电子产品类和食品与生活用品类.

3.2.3 关系模型选择

对于关系模型选择, 有两种策略:

1) 策略 1: 针对单一数据集, 符号回归方法能返回一组具有不同拟合优度 R^2 和复杂度 C 的模型, 如果模型的拟合优度和复杂度较高, 则有很大风险造成过度拟合. 用“沐浴露”这一产品种类数据进行分析, 该类产品的评论数量与购买数量的关系所对应的帕累托最优模型组中, 模型 M1, M2 和 M3 同时出现, 如图 6 所示. 由该图可知, 对于该类产品的数据集, 符号回归方法输出了一组即 9 个帕累托最优模型, 这 9 个模型均能以不同的 R^2 值解释 “沐浴露” 产品的评论数量与购买数量之间的关系, 每个模型对应的是该数据集的一个剖面.

对于最优模型的选择, 既可以选择拟合优度最高的模型, 也可以综合考虑模型复杂度和拟合优度两个指标, 并对这两个指标进行权衡. 目前有一些衡量复杂度和拟合优度的准则, 如, Akaike 信息准则(Akaike's Information Criterion, AIC)[29]、贝叶斯信息准则(Bayesian Information Criterion, BIC)[30]和 Hannan-Quinn 信息准则(Hannan-Quinn Information Criterion, HQC)[31] 等. 然而这些准则只是用来辅助决策, 对于特定领域中的特定问题, 应与专家讨论并根据领域知识制定模型的选择标准[32].

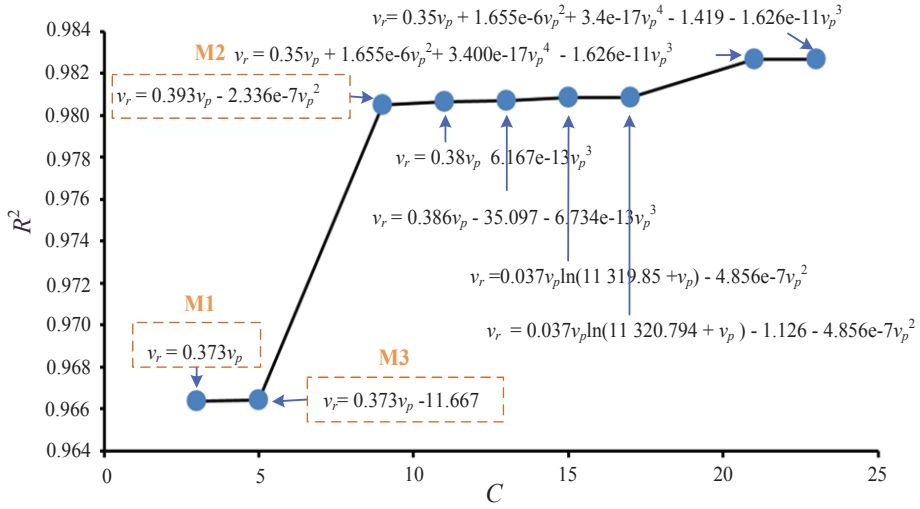


图 6 “沐浴露”产品数据的帕累托最优模型组

Fig. 6 The optimal Pareto model set for “Body Shower”

2) 策略 2: 针对多个数据集, 不同的模型有不同的平均拟合优度 $\overline{R^2}$ 和解释度 r 即适用范围. 在模型选择时, 综合考虑模型的解释度 r, 复杂度 C 和平均拟合优度 $\overline{R^2}$ 三个指标, 并对三类指标进行权衡, 具体的分两步进行. 第一步, 用复杂度和拟合优度指标构建帕累托前沿, 筛选出单一数据集中的帕累托最优模型组. 第二步, 依据适用范围和平均拟合优度指标进行排序, 在多个数据集中将模型按降序排列, 选取适用范围相对较大且平均拟合优度较高的模型, 这些模型可以作为重点推荐的结果.

如表 2 中, 模型 M2 解释了 60 % 的产品种类, 模型 M1 解释了 100 % 的产品种类, 说明 M1 的优点是适用范围广. 有 60 % 的产品种类可同时用模型 M1 和模型 M2 刻画, 说明这两个模型都能以一定的拟合优度解释 60 % 产品种类的评论数量与购买数量间的关系, 但模型 M2 的平均拟合优度更高, 也就是说, 在这 60 % 部分, M2 的优点是精度更高. 因此, 如果仅仅局限于这 60 % 的产品评论数量和购买数量的关系, 可以更倾向于采用模型 M2, 它反映的规律是: 评论数量随着购买数量的增加而增加, 但增加的速率在降低. 当然, 如果扩大到 100 % 产品种类, M1 和 M2 又是各有千秋, M1 精度略低但适用范围更广, M2 适用范围略窄但精度更高. 本文没有采用单一指标选择模型, 而是采用多个指标对模型进行综合评价, 旨在辅助人们对模型性能做出整体判断.

4　结束语

本文采用了一种数据驱动的符号回归方法, 基于大量的观测数据, 自动学习出描述产品评论数量与购买数量的关系模型和相关参数. 不仅发现了传统实证方法假设的简单线性模型, 同时还发现了新的关系模型, 如带截距的线性模型和非线性模型. 这些新的关系模型对数据有更高的拟合优度. 此外, 本文还比较分析了不同类型产品的关系模型, 分析结果显示不同产品类型下, 各模型的性能有明显的不同. 最后, 本文提出了两个模型选择策略, 可依据不同的衡量指标选择最合适的模型. 本文关于评论数量与购买数量关系的研究结果有助于商家针对不同产品类型, 合理制定评论管理策略以提高产品的评论数量.

本文的研究尚存一些不足. 目前的分类中, 每个产品类型下包含的产品种类数较少, 为了充分证明结果的可靠性, 在未来的研究中将收集更多的产品种类数据, 对不同的产品类型进行更深入和详细的分析. 同时对于产品可以按不同产品属性进行分类, 研究不同产品属性评论的数量与购买数量之间的关系. 此外, 程序在输出模型结构和参数时, 没有考虑模型系数的显著性, 下一步工作可以考虑对符号回归方法输出模型的系数进行显著性检验, 将符号回归方法与传统回归方法进行结合分析.

参考文献:

[1] Liu Y. Word-of-mouth for movies: Its dynamics and impact on box office revenue. Journal of Marketing, 2006, 70(3): 74–89.

[2] Archak N, Ghose A, Ipeirotis P G. Deriving the pricing power of product features by mining consumer reviews. Management Science, 2011, 57(8) : 1485–1509.

[3] Dellarocas C. The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. Management Science, 2003, 49(10): 1407–1424.

[4] 刘　洋, 廖貅武, 刘　莹. 在线评论对应软件及平台定价策略的影响. 系统工程学报, 2014, 29(4): 560–570.
Liu Y, Liao X W, Liu Y. The impact of online review on software and platform' s strategies. Journal of Systems Engineering, 2014, 29(4): 560–570. (in Chinese)

[5] 那日萨, 崔雪莲. 社交网络商品在线口碑情感信息传播模型研究. 系统工程学报, 2019, 34(3): 324–334.
Na R S, Cui X L. Research on online word-of-mouth sentiment information diffusion model of commercial products. Journal of Systems Engineering, 2019, 34(3): 324–334. (in Chinese)

[6] Chen Y, Xie J. Third-party product review and firm marketing strategy. Marketing Science, 2005, 24(2): 218–240.

[7] Yang X, Yang G, Wu J. Integrating rich and heterogeneous information to design a ranking system for multiple products. Decision Support Systems, 2016, 84(4): 117–133.

[8] Yang S, Hu M T, Winer R S, et al. An empirical study of word-of-mouth generation and consumption. Marketing Science, 2012, 31(6): 952–963.

[9] Duan W J, Bin G, Whinston A B. The dynamics of online word-of-mouth and product sales - an empirical investigation of the movie industry. Journal of Retailing, 2008, 84(2): 233–242.

[10] Ye Q, Law R, Gu B. The impact of online user reviews on hotel room sales. International Journal of Hospitality Management, 2009, 28(1): 180–182.

[11] Ye Q, Law R, Gu B, et al. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-mouth to hotel online bookings. Computers in Human Behavior, 2011, 27(2): 634–639.

[12] Lee J, Lee J N, Shin H. The long tail or the short tail: The category-specific impact of eWOM on sales distribution. Decision Support Systems, 2011, 51(3): 466–479.

[13] 王君珺, 闫　强. 不同热度搜索型产品的在线评论对销量影响的实证研究// 长沙: 第十五届中国管理科学年会. 2013, 21: 406–411.
Wang J J, Yan Q. An empirical study on the impact of online reviews of different product popularity on product sales // Changsha: Proceedings of the fifth Annual Meeting of China Management Science. 2013, 21: 406–411. (in Chinese)

[14] Koza J R. Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge: MIT Press, 1992.

[15] Bass F M. The future of research in marketing: Marketing science. Journal of Marketing Research, 1993, 30(1): 1–6.

[16] Anderson E W. Customer satisfaction and word of mouth. Journal of Service Research, 1998, 1(1): 5–17.

[17] Dominique M H, Leonard J P, Randall L S. Market Response Models: Econometric and Time Series Analysis, Second Edition. New York: Kluwer Academic Publishers, 2003.

[18] Vladislavleva E, Friedrich T, Neumann F, et al. Predicting the energy output of wind farms based on weather data: Important variables and their correlation. Renewable Energy, 2013, 50(6): 236–243.

[19] Khu S T, Liong S, Babovic V, et al. Genetic programming and its application in real-time runoff forecasting. Journal of the American Water Resources Association, 2001, 37(2): 439–451.

[20] Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. Science, 2009, 324(5923): 81–85.

[21] Chattopadhyay I, Kuchina A, Süel G M, et al. Inverse gillespie for inferring stochastic reaction mechanisms from intermittent samples. Proceedings of the National Academy of Sciences, 2013, 110(32): 12990–12995.

[22] Kemp C, Tenenbaum J B. The discovery of structural form. Proceedings of the National Academy of Sciences, 2008, 105(31): 10687–10692.

[23] Yang G, Sun T, Wang L, et al. Modeling the nexus between carbon dioxide emissions and economic growth. Energy Policy, 2015, 86(6): 104–117.

[24] Dellarocas C, Narayan R. A statistical measure of a population' s propensity to engage in post-experience online word-of mouth. Statistic Science, 2006, 21(2): 277–285.

[25] Thorsten H T, Gwinner K, Walsh G, et al. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet. Journal of Interactive Marketing, 2004, 18(1): 38–52.

[26] Mitchell L, Renana P, Ron S. On brands and word-of-mouth. Journal of Marketing Research, 2013, 50(4): 427–444.

[27] Sundaram D S, Mitra K, Webster C. Word-of-Mouth communications: A motivational analysis. Advances in Consumer Research, 1998, 25: 527–531.

[28] Shen W,Hu Y J, Rees U J. Competing for attention: An empirical study of online reviewers' strategic behavior. MIS Quarterly, 2015, 39(3): 683–696.

[29] Akaike H. An information criterion. Math Science, 1976, 14(153): 5–9.

[30] Schwarz G. Estimating the dimension of a model. The Annals of Statistics, 1978, 6(2): 461–464.

[31] Hannan E J, Quinn B G. The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B: Methodological, 1979, 41(2): 190–195.

[32] Yang F, Wang M. Review of systematic evaluation and improvment in the big data environment. Frontiers of Engineering Management, 2020, 7(1): 27–46.

作者简介:

杨 弦(1988—), 女, 湖南益阳人, 博士, 讲师, 研究方向: 在线评论与数据挖掘, Email: yangxian1600@126.com;

党延忠(1954—), 男, 辽宁营口人, 博士, 教授, 研究方向: 知识管理与系统工程, Email: yzhdang@dlut.edu.cn;

吴江宁(1964—), 女, 辽宁沈阳人, 博士, 教授, 研究方向: 数据挖掘与商务智能, E-mail: jnwu@dlut.edu.cn.