



北京大学

硕士研究生学位论文

题目： 某电子商务公司短生命
周期商品销售预测分析
和应用

姓 名： 刘漪多

学 号： 1201220794

院 系： 软件与微电子学院

专 业： 软件工程

研究方向： 电子商务与物流

导师姓名： 李杰 教授

二〇一五 年 七 月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

短生命周期商品的生命周期与传统商品的生命周期有着显著的差异，成熟期缩短是两者之间最主要的区别。传统的预测方法在预测短生命周期的商品时，准确度并不理想。而在电子商务环境下，商品的销量波动剧烈，增加了短生命周期商品的预测难度。

针对某电子商务企业短生命周期商品销售预测问题，提出一种新的组合预测算法，即 ESHF 算法，并使用电子商务销售真实数据对 ESHF 算法进行了验证。实验结果证明，ESHF 算法的准确度高于传统的预测算法 DES 和 ARIMA 算法，也优于组合算法 S-ESHF 算法。

ESHF 算法实现了小规模商品的批量预测，帮助业务人员节省了时间。

本文的应用创新价值主要体现在以下几个方面：

第一：提出在电子商务环境下，商品销量的异常数据均有其现实的意义，因此采用七点均值法处理离群点，以保留商品趋势特征的同时降低商品销量的过度波动给预测带来的难度。

第二：通过研究目前的预测算法，发现使用分段三次样条 Hermit 插值法替换 EEMD 中的三次样条插值法，可以降低因数据波动过快而产生的过冲和欠冲现象。并结合 ARIMA 预测算法，组合成为本文的应用算法 ESHF 算法。

第三：使用企业中的真实数据对 ESHF 算法的准确度进行验证，证明算法预测结果的准确度高，并可以在企业中应用，解决企业面临的真实存在的问题，具有较高应用价值。

第四：结合企业中商品的库存维度数据，提出预测销量-库存四象限模型，将商品分成四个类型，并对不同类型的商品提出了具体的业务操作，进一步提高了业务人员的工作效率。

关键词：电子商务，短生命周期，EMD，EEMD，ARIMA 预测

Sales forecasting analysis and application of Short-Life-Cycle products in a E-commerce company

Yiduo Liu (Software Engineer)

Directed by Jie Li

ABSTRACT

Distinct from the traditional products, the short-life-cycle products is noted with major differences in the reduced maturity stage in its lifecycle, and the imprecise results as using traditional forecasting theory and methods to forecast its sales did not turn out to be satisfactory. In today's e-commerce environment, sales of products has become far more variable, which also makes it far more challenging to forecast the sales of short-life-cycle products with precision.

In order to solve the sales forecasting problem of the short-life-cycle products in the e-commerce company, a new combined forecasting algorithm, i.e. ESHF Algorithm, is introduced and verified with real data in business. Results of the experiments show that ESHF exceeds in precision as compared with traditional forecasting algorithms of DES, ARIMA, and also the combined algorithm S-ESHF. ESHF can be cost-saving and time-convenient in sales forecasting in batches for small-scale products

The innovation values of this paper are reflected in the following aspects:

- I. The deviated data of product sales in e-commerce environment are analyzed to figure out the realistic reasons lying beyond. In order to reducing the variable in product sales while reflecting its trend, seven-point-average-mean-value method is applied in processing the outlier points.
- II. With research over current forecasting algorithms, it is noted that the overshoot and undershoot distortion from data variation can be reduced by replacing the Cubic Spline Interpolation of EEMD with Hermit Spline Interpolation. Based on this improved method and the forecasting algorithm ARIMA, ESHF is established and introduced.

- III. The precision of ESHF is verified as using real data in business, which proves the practical utility of applying ESHF in enterprise. Furthermore, Sales-Inventory Quadrant Module is introduced as based on the product inventory data in enterprise, which can provide specific recommendation over the sales forecast of different types of products, and also help improving productivity in enterprise.

KEY WORDS: E-commerce, Short-Life-Cycle, EMD, EEMD, Forecasting of ARIMA

目录

第一章 绪论	1
1.1 选题背景和意义	1
1.2 研究目标和内容	2
1.3 论文结构和组织	3
第二章 相关概念及国内外文献综述	7
2.1 相关概念	7
2.1.1 生命周期	7
2.1.2 短生命周期	8
2.1.3 电子商务环境下的商品生命周期特征	8
2.2 国内外研究现状	9
2.2.1 国内研究现状	9
2.2.2 国外研究现状及算法综述	11
第三章 相关分解算法和预测算法	15
3.1 分解算法	15
3.1.1 EMD 算法	15
3.1.2 EEMD 算法	18
3.2 预测算法	20
3.2.1 ARIMR 模型	20
第四章 电子商务环境下短生命周期商品销量预测方法	23
4.1 ESHF 算法的预测流程	23
4.2 数据清洗	24
4.3 HEEMD 算法	26
4.4 ESHF 算法流程	28
4.5 ESHF 算法准确度的评价体系构建	30
第五章 实验设计和应用效果验证	31
5.1 公司简介和问题描述	31
5.1.1 公司介绍	31
5.1.2 问题描述	32
5.2 实验设计和算法测试	34
5.2.1 实验环境设计	34

5.2.2 数据收集和预处理.....	34
5.2.3 ESHF 算法应用实例	37
5.2.4 大规模数据测试.....	41
5.2.5 四种算法预测结果分析	43
5.3 商品的分类和业务举措.....	49
5.3.1 预测销量-库存四象限模型应用	50
5.4 ESHF 算法评价与展望	52
第六章 结论和展望.....	53
参考文献.....	55
致谢.....	59
北京大学学位论文原创性声明和使用授权说明.....	61

第一章 绪论

1.1 选题背景和意义

近年来,中国的电子商务发展的速度越来越快,规模越来越大,电子商务环境对商品的销售有着巨大的影响,线上市场的影响举重若轻(本文中线上市场指通过电子商务网站进行商品的售卖行为,包括B2B, B2C和C2C等模式)。电子商务已经进入规模发展阶段,交易额从2009年的3.7万亿元增长到了2013年的10.7万亿元。甚至在2013年的双十一,仅天猫和淘宝当天的总交易额就达到350.19亿。根据中国情报网(<http://www.askci.com/chanye>

/2015/01/22/10486tdaw.shtml)的数据,截止到2014年,电子商务交易额已经占到零售业的49.6%。可见电子商务的发展迅猛,使得线上市场的地位越来越高。

相较于线下市场,线上市场的竞争更为激烈,但也为线上市场中的企业造就了更大的机遇和更多的挑战。线上的物品丰富,操作简单,配合发达的物流网络,上网成本降低,种种因素使得人们越来越愿意在线上购物。线上的市场和线下的市场存在着显著的差异,一方面,线上市场可以获得用户最新的消费冲动点,迅速反应,抓住商机;另一方面,线上市场的竞争强度远高于线下市场^[2]。线上购物的便捷,给消费者提供更多的选择的同时,降低了消费者的忠诚度。面对这些机遇和挑战,需要线上市场中的企业在抓住商机的同时,必须快速做出与竞争对手差异化的反应,才能抓住消费者的眼球。

电子商务环境下,企业的盈利点更多的在于供应链管理和线上促销活动的管理。线上商品价格透明,企业的议价空间缩小,商品的利润不能再跟线下市场一样,直接通过商品的定价权来控制利润的空间。线上企业想要获得更高的利润,需要更加关注整个供应链的其他环节。传统的供应链是运输和物流学的延伸,而电子商务环境下的供应链可以看成三个方面,一是电商企业和消费者之间的联系,涉及消费者的线上购物行为,公司的促销活动设计,信息反馈和售后服务等;二是企业和企业之间的联系,例如,企业之间的系统的对接,进销存信息的协同管理等操作;三是企业内部之间的协调,包括网站和数据的底层架构对整个公司的应用系统的支撑等^[3]。

本文假设企业内部的协调和企业间的合作良好,没有问题。那么线上的供应链就可以简化为线上活动策划和线下库存盘货配合的过程。

所谓促销活动就是,在线上为某些或某一品类商品搭建一个页面,并对活动中的商品进行打折或直降处理,从而吸引更多的消费者前来网站购物。而线上销

售主要也是以活动的方式进行，活动会带来更多的流量，流量就是消费者在网站浏览路径，对于电子商务网站而言，没有流量就没有点击，没有点击就不会有订单的产生。而流量增多最直接的结果就是会增加商品的曝光，也就是给商品带来更多的展示机会，让更多的消费者有机会进一步的了解商品，才有可能快速提升商品的销量。而一个商品如果想要上一个活动，就必须保证这个商品的库存是充足的，没有质量问题的。这样才能保证活动的转化率，转化率是指订单量和流量的比值，转化率高则证明活动带来的订单量高，也就是活动效果好，反之转化率低则代表订单量少，流量没有转化成订单，造成了流量的浪费。但是，如果为了活动而过多的采购商品，则会造成商品库存的积压，一旦商品积压滞销，企业为提高库存周转，减少损失，最终只能采取倾销的手段以低于进价的价格出售商品，这样以损失商品毛利为代价来换取更多的库存空间的行为，最终给企业造成了损失。为了更好做到线上活动和线下库存的良好配合，就需要尽可能准确的提前预知商品的销量，以帮助指导企业业务人员提前精准备货。

在这样背景下，准确的预测商品的销量，可以有效的指导备货，降低企业损失。但是因为线上线下市场的商品销售的特征差异过大，原本在线下所使用的销售预测方法并不完全适用于线上的市场环境。本文就电子商务环境下的商品销售预测问题进行预测算法的探索和算法准确性的评估，并展示在实际业务中，算法的应用场景。

由于线上面临的市场范围远远大于线下市场，不能把商品线下的销售情况完全的移植到线上。由美国Frank Bass提出的BASS模型，在最开始应用于耐用品的销量预测，预测效果良好。BASS模型对市场需求相对稳定的商品的销售预测效果也比较好，但BASS模型需要考虑商品的历史数据，通过历史数据来推测市场需求，因此历史数据越多预测的准确性越高。而线上很多商品，诸如生鲜、食品、电子设备等商品，它们或短期应季销售，或更新换代过快，这些商品的生命周期都在逐渐缩短，且不能够提供充足的历史数据以供于预测市场需求。因此一些诸如BASS模型等传统的预测算法在线上环境的商品销量的预测准确性很低，不足以满足指导企业备货的需求。因此，本文研究短生命周期商品在电子商务环境下的销售情况并对下一期销量进行预测，预测结果约准确越有利于电商企业控制成本，具有理论研究和现实的双重意义^[4]。

1.2 研究目标和内容

正如在选题背景中所介绍的，目前的预测算法多应用于具有长期的历史销售数据的商品，或应用于长生命周期商品的销量预测。在预测短生命周期的商品时，具有一定的局限性，因此找到合适的算法来预测短生命周期商品的销量，有利于

指导企业进行合理备货。

短生命周期商品很多，比如电子产品，像手机，电脑，照相机等商品，都是短生命周期的商品。而手机是一个典型的短生命周期的电子产品，尤其是在智能机兴起后，人们越来越关心手机的各种软硬件配置和性能，比如屏幕大小、分辨率、电池待机时间、手机的运行速度，反应速度，甚至手机平台也是用户选择的指标之一。而随着手机的版本快速迭代更新，用户更换手机的频率也越来越快。

在预测像手机这样的短生命周期商品的销量时，要特别考虑商品的生命周期短这一特征，这类商品的历史数据通常具有不稳定性与非线性的特征，为准确预测商品销量带来一定的难度。如何克服数据的不稳定性与非线性为算法带来的影响，准确识别商品的生命周期，结合商品的背身的特征对商品的销量进行综合预测，以找到能够准确预测电子商务环境下，短生命周期商品销量的算法是本文的研究重点之一。

本文将根据在某电子商务公司搜集而来的短生命周期商品的历史销售数据进行预测分析，采用以 EMD 算法为基础的改进算法进行数据分解处理工作，再使用 ARIMA 算法来预测其下一期的销售量，指导企业在合理备货的同时，也为线上活动保证合理的库存，实现了线上活动和线下库存之间的良好的配合。

1.3 论文结构和组织

根据研究情况，本文共分为六部分。

第一部分为绪论，主要说明选题原因，并简要介绍行业背景，同时说明研究的目标和意义所在。电子商务发展迅猛，电子商务环境下诸多商品的生命周期逐渐压缩成为短生命周期的商品，且商品的历史销量数据较短，波动较大，难以凭借肉眼或传统算法来发现其中的规律以准确预测出下一期的商品销量情况。在这样的背景下，研究能够准确预测电子商务环境下短生命周期商品的算法是具有理论研究和现实指导的双重意义的。

第二部分是国内外的研究现状，主要介绍国内外在相关领域的研究情况和成果，并通过深入分析研究目前流行的算法，综合比较算法的应用条件和预测结果的准确度，为本文的组合算法的研究提供理论基础。

第三、四部分是本文的核心算法和相关算法的介绍。第四部分介绍的是本文最终采用的核心组合预测算法的逻辑和预测结果准确度的评价体系，是从第三章算法的基础上改进演变而来，本文安排在这两章中详细介绍相关算法。

第五部分为本文的核心算法应用部分，也是本文的核心部分。本文通过描述企业在现实中存在的真实问题，并收集企业中的真实数据以验证算法的准确性和合理性，同时收集企业中的其他维度数据，结合商品的预测结果，针对不同商品

给出解决库存和销量问题的方案。

第六部分为本文的结论与展望。在本章节，总结本文研究和应用成果并进行展望。本文找到能够较为准确的预测电子商务环境下，短生命周期的算法，并用于解决现实问题。但在如何在商品多样化的现实应用中进行推广，还需根据企业自身情况进行改进。

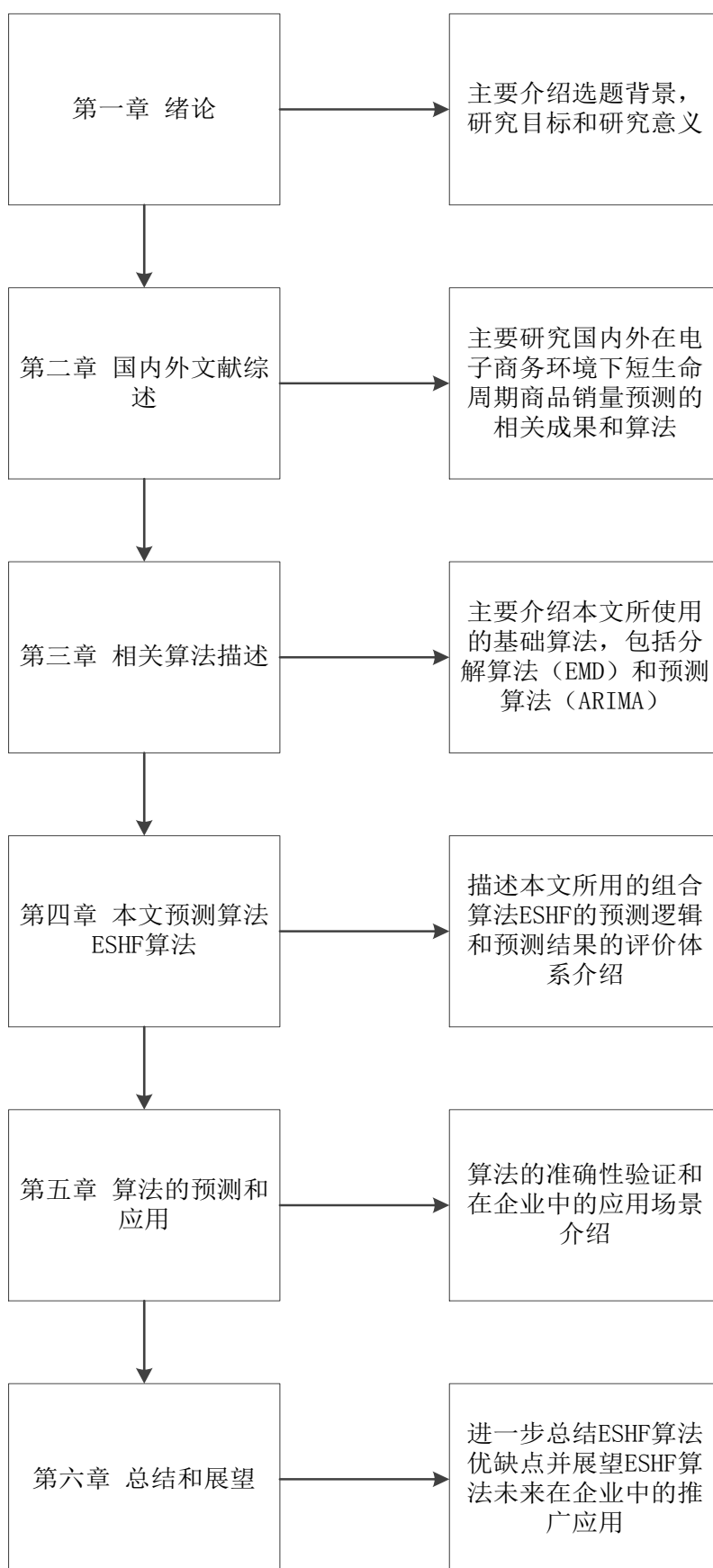


图 1.1 论文架构图

第二章 相关概念及国内外文献综述

2.1 相关概念

2.1.1 生命周期

生命周期的概念有着非常广泛的应用。在生物学中，一个生物的生老病死便是一个生命的周期。而在机械工业领域，一个商品从生产到最后损毁退出市场，便是一个商品的生命周期。一个商品的生命周期的一般认为是一条钟形曲线，如图 2.1，一般分为四个阶段，分别为：引入期、成长期、成熟期和衰退期。这里，商品的生命周期不是指它的使用周期，而是它从进入市场到最后退出市场的周期[5]。

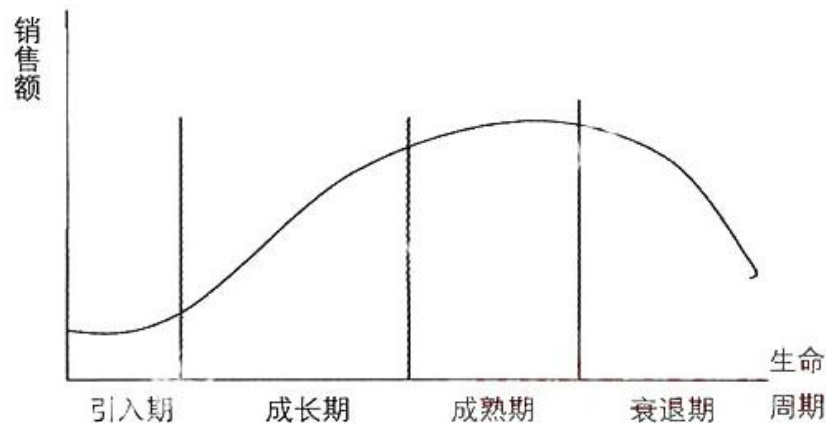


图 2. 1 商品的生命周期图示

有效的识别生命周期在预测商品销量时，具有非常重要的意义。商品处于生命周期的不同状态时，市场对它的需求是不一样的，它的增长趋势有着显著的差异。当商品刚刚进入市场处于引入期时，消费者对于该商品的认知较少，不会有大规模的购买行为，它的增长速度缓慢。而随着企业慢慢的推广该商品，使得商品得到消费者的认同，推动商品进入成长期，商品处于成长期时，它的销量增长速度会大大加快。当商品进入大规模生产期时，市场需求接近饱和，商品销售量会处于一个相对稳定的状态，销售量不会产生大幅度的变动。而当商品的替代品和新一代的商品进入市场后，该商品会进入衰退期，慢慢的退出市场。处于衰退期的商品的销售量会慢慢下滑。因此，识别商品在生命周期中所处的阶段，可以提高预测下一期商品销量的准确度。

对于识别商品的生命周期，可以基于 logistics 函数采用聚类的方法来识别，预测商品未来的销量。杨俊、余安乐等人在 2012 年提出，采用 BASS 模型和指

数平滑法，在系统动力学的基础上，对生命周期具有良好的识别和预测效果^[6]。

2.1.2 短生命周期

短生命周期商品是指销售周期较短的商品，比如季节性商品。随着科技和经济的发展，人们的消费需求从单一化变为多元化。为了更好地满足消费者的需求，厂商通过提高技术，压缩生产商品的周期，通过提高创新能力，加快商品的更新速度。商品的生命周期变短，已经是一种不可避免的趋势。电子商品的变化尤其显著。

短生命周期商品具有以下特征：

① 商品的生命周期短暂，正如上文所说，企业为了保证市场地位，抓住顾客眼球，摆脱模仿者，会加快商品的更新换代速度，不断创新出新的商品，保持商品线的活力；

② 传统商品的商品生命周期相比，短生命周期的商品的成熟期较短。如果商品具有较短的成熟期，在预测时，有可能识别不出进入成熟期的拐点，从而错失商机，因此短生命周期商品加大了企业预测商品未来销量的难度；

③ 具有高可替代性。短生命周期商品多为价格波动过大的商品。这类商品（如手机）一般在功能上和样式上具有完全的可替代性。新的竞争商品进入市场时，往往可以替代现有商品的所有功能并加以创新。因此新商品的进入，会取代老商品，促使老商品的价格下调，同时影响老商品的销售状况^[4]。

2.1.3 电子商务环境下的商品生命周期特征

传统市场下商品的生命周期不管长短，历史的销量情况是相对稳定的。而电子商务环境下，商品的销售情况极易受到不可控的因素影响，造成某种商品销量的暴涨或暴跌，当下的销量数据成为噪声数据，在使用历史的销量数据进行预测时，这类数据会影响预测结果的准确性。

比如，2014年某位微博名人在微博上发了一条状态，抱怨某电子商务公司送货慢。这个商品是价值200元的电脑桌。此微博一出，网友们纷纷上该电商网站搜索该商品。该商品在这种不可控制的因素下，销量突涨。类似的事情还有某明星用遥控飞机求婚，使得同款玩具飞机的销量突然上涨。这种名人效应在电商平台上对商品的销量影响非常显著，但是这种影响不具备可持续性，商品的销量会慢慢的回归正常值。但当时的销量情况作为历史数据保留了下来，这并不利于商品生命周期状态的识别，影响销量预测的准确性。对于这样激增的数据，需要在进行预测之前就进行数据的预处理。本文在使用组合预测算法进行预测之前，先对这类激增的销量数据进行了预处理，降低异常数据对预测结果的准确性造成的不良影响。

2.2 国内外研究现状

本文在研究算法的同时，用企业的真实数据和所面对的问题来检验组合算法的准确性和实际可操作性。考虑到国内外的国情差异，在算法应用层面更多的参考国内文献。而在算法的解析和改进方面，主要参考国外文献。

在本节，研究国内外在短生命周期预测算法在各个领域内的研究成果，根据在各个领域内发表的文献数量的统计，当前少有人在电子商务环境下进行相关的理论研究，可以看出本文的研究方向在一定程度上填补了这段空白。另外，在国内外文献综述部分，本文对现有的一些比较流行的算法做了简单比较分析。

2.2.1 国内研究现状

从中国知网（www.cnki.net）上统计 2000-2014 年，短生命周期预测的相关文献，我们发现，近 14 年共发表相关的期刊文献 814 篇，呈上升趋势，见图 2.2；这些文献主要涉及企业经济和宏观经济与可持续发展等四十多的个领域，见图 2.3。根据文献数量及研究学科领域，我们可以看出，短生命周期的预测算法可以应用于多种领域，对短生命周期商品的预测在多个产业领域有应用场景。。

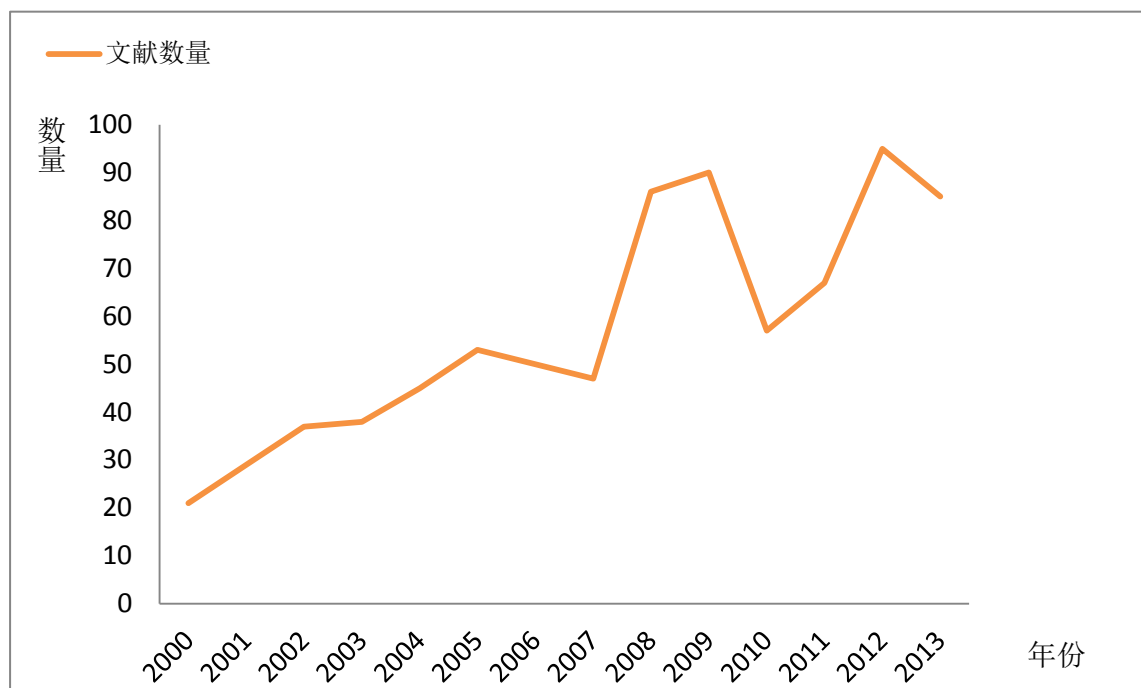


图 2. 2 近 14 年短生命周期预测相关文献发表数量统计

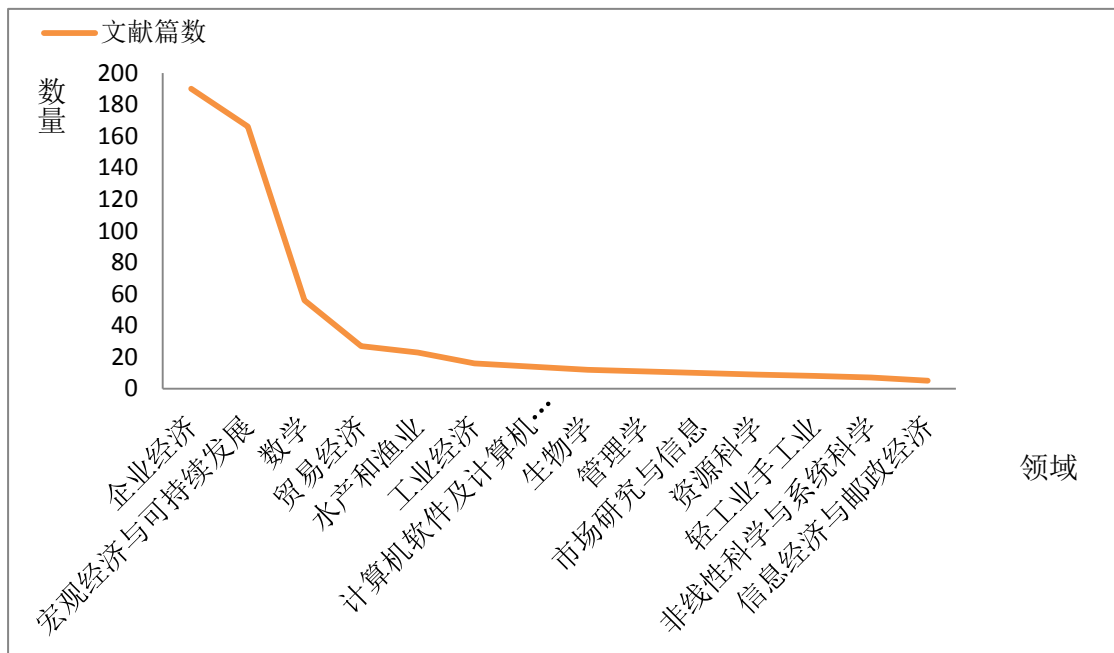


图 2.3 近 14 年短生命周期预测文献在各个领域内的分布情况

图 2.3 仅统计了文献篇数在四篇以上的领域，而在电子商务领域进行预测的相关研究文献较少，仅有两篇，分别应用于旅游业^[7]和客户流失预测^[8]。没有应用于电子商务环境下短生命周期商品的预测的算法。不考虑应用环境，仅考虑短生命周期商品预测的算法，总结如下：

在供应链管理方面，由于线上线下的销售渠道的差别，两种渠道的需求量差异较大。对于短生命周期商品而言，为了更好的预测两个渠道的需求总量，使用 Bass 方法和 SVM 方法进行组合预测^[9]，SVM 是支持向量机算法的简称，由 Corinna 等人在 1995 年提出。使用 Bass 算法和 SVM 方法并结合其他机器学习方法可以在多个领域内解决现实问题。Bass 模型更加适用于处理小样本的数据，而 SVM 算法适用于对多维度的数据进行预测。考虑到季节因素、顾客数等相关因素，构建需求预测模型，分别预测线下线上的需求量。实验证明，该模型的预测准确性较高。

刘峥、徐琪等人提出，在原有的 Bass 模型基础上，通过服务因素和商品复购率等因素，改进了 Bass 模型，并将改进的模型应用于服装行业，取得了较好的效果^[10]。周凯歌、徐琪等人同样改进了 Bass 模型，在 Bass 模型的基础上，加入了商品生命周期和季节性影响因子，并通过实际案例，证明了模型的可用性^[11]。

徐贤浩、陈雯等人在研究商品订货策略时，考虑到实际商品库存因子和需求因子，商品生命周期的时间短等相关问题，建立了三种订货策略。通过求解订货策略，分析对订货量、库存成本等因子对模型的影响，结合 Bass 模型函数，改进订货策略。并应用于实际生活中，指导企业在不同的时期备货数量应有差别^[12]。

徐贤浩、郭晓云在研究短生命周期预测算法时,使用了 Norton 模型。结合短生命周期的特征,并采用系统仿真算法验证了 Norton 算法在短生命周期预测算法中的可行性,最终利用 MP3 的历史销量数据证明 Norton 算法的预测算法比较好^[13]。但是文中使用的数据跨度有两年,而现在一般认为,短生命周期的商品的全生命周期最多为 18 个月。该算法所需的商品生命周期过长,不适用于电子商务环境下,生命周期短且销量快速变化的商品。

徐贤浩、宋奇志对短生命周期商品销量预测算法进行了比较研究,发现短生命周期商品的预测多为新品预测或某类单一商品的预测,如个人电脑等。在预测方法上,多使用统计学方法或扩散模型。然后统计学方法在实际问题中的应用较少,不能为短生命周期商品的预测提供直接的通用的办法;而扩散性模型多需要专家经验的支持^[14]。

总结以上文献算法即应用场景,可以发现在短生命周期预测的领域内,多应用 Bass 模型及其改进模型。这些模型对商品生命周期的识别有较高的要求。然后,对于短生命周期的商品而言,由于其生命周期与传统生命周期相比,具有的不确定性,且导入期长,成熟期短,不能直接沿用传统商品生命周期的识别办法。而扩散型模型,需要专家的经验支持。而在当今电子商务的环境下,商品生命周期变化各异,相关影响因素过多,领域内的专家经验不足以支撑多品类的预测和生命周期的识别。因此,本文研究新算法应用于电子商务环境下的短生命周期商品的预测是具有现实意义的。

2.2.2 国外研究现状及算法综述

分别按照关键词 sale forecasting、life cycle forecasting 和 e-business forecasting 搜索美国 ACM 数据库(Association for Computing Machinery, 美国计算机学会)和 ScienceDirect 数据库,经手工去重后,共收集到文献 155 篇,其中来自 ACM 98 篇, ScienceDirect 57 篇;从文献在关键词中的分布情况,见图 2.4,可以看出,国外在销售预测和电子商务预测的研究成果远高于生命周期预测的成果。而在生命周期预测的领域内,研究成果较少。

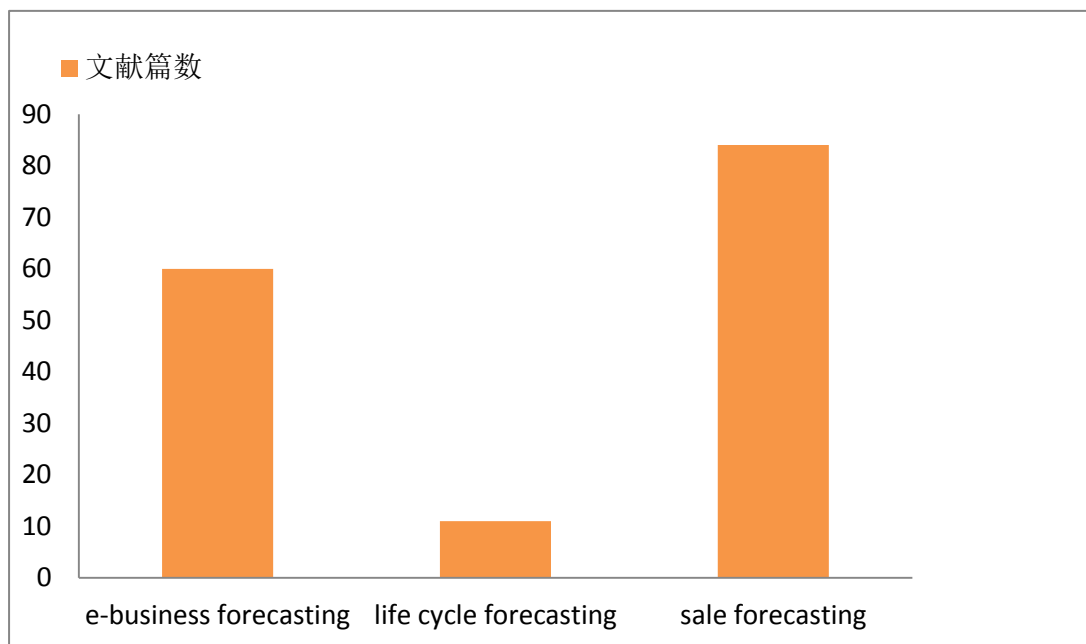


图 2. 4 国外相关文献在各个关键词上的分布

收集各领域内相关算法，主要总结为 6 类。分别为 1) 增长模型；2) 模拟法；3) 时间序列法；4) 分解法；5) 数据挖掘法和 6) 其他方法。下面主要介绍下这六类算法：

1) 增长模型

Xu 和 Song 分析了短生命周期商品的特征，并改进的 Bass 模型，用于短生命周期商品的预测^[15]。他们考虑到了需求季节性因素，并在 Bass 模型中添加了季节因子。和原始的 Bass 模型相比，在预测短生命周期商品时，这种改进的 Bass 模型的效果更好。但是，这个 Bass 模型需要使用所有的历史数据来构建，而在实际问题中，有许多商品处于引入期，没有足够的历史数据来训练模型。在这种情况下，当预测下一期的需求时，使用历史数据构建了 Bass 模型，但是在商品进入成熟期之前，很难估计全部的潜在市场需求，导致模型的预测效果不好。因此，增长模型很难用于预测市场需求。正如 Xu 和 Zhang 等人所说，Bass 模型更加适用于预测那些特定的商品型号或者特定的产业的需求^[16]，而不适用于预测短生命周期的商品的需求。

2) 模拟算法

2010 年 Szoza 提出一种新方法，即使用与正处于引入期的商品相似的商品的生命周期来预测该商品的销量^[17]。Green 和 Armstrong 提出了结构化类比法，在预测某新商品前，先通过采访专家，尽可能多的找到与目标商品类似的商品，通过判断这些商品的结构来预测新商品的销量^[18]。对于拥有相似商品的新商品，这种方法的效果很好，但是在电子商务的环境下，即使是互相可替代的商品，在

环境不同的情况下,历史的销售趋势也可能大不相同。尤其是短生命周期的商品,很难找到足够多的类似的商品,通过这些商品的生命周期来预测新商品。Wu 等人在 2006 年探索比较了多种领先指标的方法,用与短生命周期商品的预测。通过这些领先指标和商品的之间的相关关系,可以提前预测出 2-8 个月的需求模式^[19]。可是在电子商务环境下,一些特定的商品,例如为平台差异化销售,指厂家会为某一个电商平台专门设计生产的商品。该商品仅在某一个平台上销售,这种商品称为平台专供品,专供品是很难找到领先因素的,这时方法就失效了。

3) 时间序列算法

时间序列方法已经被广泛的应用于需求预测领域。例如,Johnson、Thompson 和 Ray 都先后在库存控制等问题上使用了 ARIMA 模型^[20, 21]。Miller 使用指数平滑法构建了模型化商品需求^[22]。然而,当数据是非线性和不稳定时,时间序列预测的方法的准确性会大大降低。

自回归积分滑动平均模型,即 ARIMA (Autoregressive integrated moving average model)模型。ARIMA 模型在经济学科和统计学科有着非常广泛的应用。ARIMA 模型最初由 Box 和 Jenkins 提出,是一种非常流行的时间序列预测的方法^[23]。ARIMA 的模型函数为 $ARIMA(p,d,q)$, P 代表自回归项, q 代表移动平均项, d 代表差分项。ARIMA 模型可以将不稳定状态的时间序列数据转变为稳定状态。因此,通常认为 ARIMA 模型是 ARMA 模型的拓展。

4) 分解算法

2012 年, Chung 等人在 Blockbuster 公司提出了电影和游戏的销售预测的模型^[24]。他们假设销售由三部分组成:

- ① 已经有过购买行为的消费者;
- ② 准备购买的消费者;
- ③ 以上两类消费者间的网络影响关系。

经验证,他们通过模型的预测出的销售结果和实际发生的销售行为具有很高的匹配程度。也就是说,这个销售模型预测的准确性非常好。

Hilbert-Huang Transform(HHT)算法由 Huang 等人分别在 1996、1998 和 1999 年提出并加以改进^[25-27],这种算法对具有非线性和高不稳定性的数据建模效果非常好。EMD (Empirical Method Decomposition, 经验模式分解)算法,是 HHT (Hilbert-Huang Transform)算法的第一步,这种算法以数据本身建立后置检验模型,将数据进行分解,分解出有限个 IMF (Intrinsic Mode Function, 本征模函数)^[28]。EMD 算法是基于数据本身,不依赖与任何其他算法,非常适用于非线性、高不稳定状态的数据。它可以将这样状态下的数据分解出一系列带着数据本身的波动情况的 IMF。比较 EMD 其他算法,如小波分解和短时傅里叶变换等算

法, EMD 的在处理非线性、不稳定的数据时, 具有更好的效果。

尽管如此, EMD 算法还是有些弱点的^[29]。本文将在 EMD 的基础上, 加以改进, 使用新算法进行电子商务环境下的短生命周期商品的预测。

5) 数据挖掘算法

Maaß 等人讨论了如何使用数据挖掘技术来提高短期预测算法来预测短生命周期商品^[30]。他们研究发现, 数据预处理对预测结果的准确性有着至关重要的影响。Shearer 在 2000 年探索出, 数据预处理包含五个要素, 分别是: ①数据选择; ②数据清洗, 清除噪声数据; ③重构数据; ④整合数据; ⑤标准化数据^[31]。

本文算法中, 我们先对数据进行清洗, 然后进行分解, 寻找到数据的特征。之后再对数据进行预测。具体算法过程, 会在第四章进行详细的描述。

6) 其他方法

Zhu 和 Thonemann 提出适应性预测算法, 使用结构知识来预测需求, 并用于探索最优库存策略。众所周知, 最优库存策略是非常难以选择的, 于是他们进一步提出了三种库存启发式算法以找到最优库存策略, 结果显示, 其中一种算法可以接近于最优库存策略^[32], 同时他们使用北美的一家个人电脑制造商的数据来检验他们的算法, 结果证明该算法的准确性较高。

Cakanyildirim 和 Roundy 研究了许多预测需求的算法, 发现这些算法都很容易受到随机误差的影响^[33]。他们提出了一种方法, 通过计算预测误差相关性和平方差, 并通过这个方法来调整预测结果

第三章 相关分解算法和预测算法

本章主要从理论上介绍分解算法 EMD (Empirical Mode Decomposition, 经验模式分解) 算法和 EEMD (Ensemble Empirical Mode Decomposition, 集合经验模式分解) 算法, 预测算法 ARIMA (Autoregressive integrated moving average model, 自回归积分滑动平均模型), 并分析这三个算法分别在电子商务环境下短生命周期商品预测上的应用和局限性。

3.1 分解算法

由于电子商务环境下短生命周期商品的销量具有不稳定性、历史销量数据少等特征, 降低了已有预测算法的准确性, 例如 DES 算法 (Double Exponential Smoothing, 双指数平滑法) 和 ARIMA 算法等。为了提高算法的准确度, 需要先将数据分解成多个线性稳定的分数据, 对分别对这些分数据进行预测, 再将预测结果叠加, 成为最终的预测结果。本文使用的分解算法改进 EEMD 算法, 而 EEMD 算法是 EMD 的改进算法。故而本节将主要介绍的分解算法为 EMD 算法和它的改进算法 EEMD 算法

3.1.1 EMD 算法

正如本文第二章介绍, Huang 等人在上世纪 90 年代提出的 HHT (Hilbert-Huang Transform) 算法可以得到不稳定非线性信号的 Hilbert 谱。而 HHT 算法的第一步正是 EMD 分解算法。简要概括 HHT 算法的过程如下: 先通过 EMD 算法识别信号中的瞬时特征, 并将这些瞬时特征从不稳定的原始信号中分解出来, 再对这些分解出来的平稳线性的信号采取 Hilbert 变换, 最后将这些变换后的信号进行叠加, 便得到了原始信号的 Hilbert 谱。在整个算法的过程中, EMD 算法的作用就是将一个不稳定的原始信号分解成多个稳定的信号, 再进行后续的合并处理^[34]。

不稳定信号其实是由不同尺度的波互相叠加, 互相影响而形成的。对这种不稳定非线性的信号进行 EMD 分解, 可以将这个内在的影响信号波动的细节信号分解出来。产生的有限个稳定的线性函数, 也就是本固有函数或本征模函数, 即 IMF, Intrinsic Mode Function。

IMF 必须满足两个条件:

第一：在整个信号的全过程中，穿过极值点（极大值或者极小值均可）的次数和穿过零点的次数之差最大为 1。假设穿过零点的次数为 N ，穿过极值点的次数为 M ，则 $|N-M| \leq 1$ ；第二：在整个信号过程中，任意一点所对应的由极大值点通过三次样条函数构建出来的上包络线和由极小值点通过三次样条函数构建出来的下包络线上的点的均值的差为 0。即，假设原始信号上任意一点为 $O(t)$ ，所对应的上包络线和下包络线的点分别为 $U(t)$ 和 $D(t)$ ，则需满足

$$O(t) - \frac{U(t) + D(t)}{2} = 0 \quad (\text{公式 3.1})$$

当信号满足以下条件之一时，需要进行 EMD 操作，以得到平稳的信号来进行后续操作^[35]。

① 假如原始信号的零点个数比极值点数目至少少两个；即假设原始信号有 A 个零点， B 个极值点，则需满足 $B-A \geq 2$ 。

② 在整个信号过程中，任意一点所对应的由极大值点通过三次样条函数构建出来的上包络线和由极小值点通过三次样条函数构建出来的下包络线上的点的均值的差不为 0。即，假设原始信号上任意一点为 $O(t)$ ，所对应的上包络线和下包络线的点分别为 $U(t)$ 和 $D(t)$ ，则满足

$$O(t) - \frac{U(t) + D(t)}{2} \neq 0 \quad (\text{公式 3.2})$$

EMD 分解的步骤为：

1. 找到原始信号 $X(t)$ 的上下包络线。通过找到信号中所有的极值点，包括极大值和极小值，分别对极大值和极小值使用三次样条差值法得到上包络线和下包络线；找到上下包络线的均值线 $M(t)$ 。

2. 判断 $H(t) = X(t) - M(t)$ 是否满足 IMF 的特征。如果 $H(t)$ 满足 IMF 的特征，则 $H(t)$ 是信号 $X(t)$ 的第一个 IMF 分量，如果不满足，则用 $H(t)$ 替代 $X(t)$ ，重复 1、2 操作，直至 $H(t)$ 满足 IMF 的条件；

3. 当找到第一个 IMF 后（假设为 C_1 ），则差分数据序列变为 $R_1 = X(t) - C_1$ 。

4. 重复 1、2、3 过程， R_n 不可分。最后就找到了所有的 IMF 和 $X(t)$ 的均值或趋势 R_n ^[35]。

整个 EMD 的过程用图形表示为：

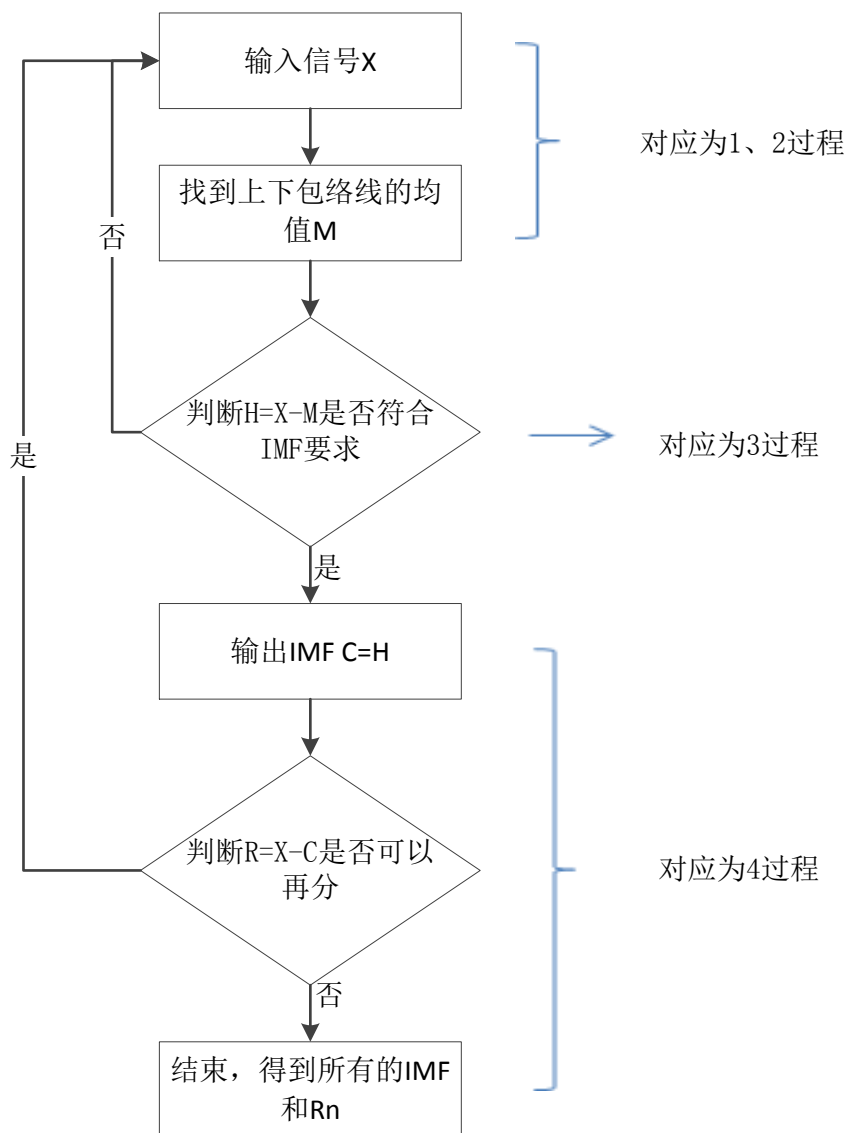


图 3. 1 EMD 分解过程流程图

虽然 EMD 算法对信号有着良好的自适应性，在分解过程中，可以保持信号本身的特性，对于信号降噪具有非常好的效果^[36]，但是 EMD 算法仍然存在着一些问题。如上所说，EMD 分解过程其实就是对信号不断的进行过滤，过滤出符合条件的信号。但是在过滤的过程中，信号在两端处是发散的，随着过滤的推进，发散的会慢慢向内推进，影响最终得到的 IMF 效果，使得 IMF 失真^[34]。而对于快速变化的信号，可能会出现信号间复杂多次叠加，使用 EMD 对于快速变化的信号进行分解时，最终得到的 IMF 容易出现模式上的混淆^[37]。而短生命周期的商品的波动变化相较于传统商品的波动更快速，而在电子商务环境中，不仅加快了短生命周期商品的波动，还会频繁出现异常事件，影响短生命周期商品的历史数据，如前文提到的互联网的不可控的

网络外部影响因素，诸如微博事件和名人效应等。因此，不能直接使用 EMD 在电子商务环境下进行短生命周期商品销量的分解预测。

3.1.2 EEMD 算法

正因为 EMD 算法还有些缺点，Wu 和 Huang 在 2009 修正了 EMD 算法。他们提出对原始信号增加白噪声数据，通过这些白噪声数据，使得原始数据在各个时瞬间上变得连续^[38]。EEMD 通过给原始信号增加在统计学上具有均匀稳定特性的高斯白噪声数据，可以有效的解决由于不同尺度的信号间，极值点分布过多或过少或分布不均，所造成的 IMF 的模式混淆问题，同时也在一定程度上改善了 EMD 的端点问题^[39, 40]。

EEMD 算法的步骤为：

1. 像原始信号 $X(t)$ 中加入高斯白噪声数据 $N_i(t)$ ，得到新的信号 $X_i(t)$ ；

$$X_i(t) = N_i(t) + X(t) \quad (\text{公式 3.3})$$

2. 对信号 $X_i(t)$ 进行 EMD 分解，得到 IMFS_i 族 C_i ；
3. 多次重复 1、2 步骤，每次加入不同的高斯白噪声数据，重复的次数根据实际信号/数据情况来确定；
4. 将得到的所有 IMF 的均值作为最后的结果 C ^[39]。C 的表达式为：

$$C = \frac{1}{N} \sum_i^N C \quad (\text{公式 3.4})$$

EEMD 算法的图示为图 3.2：

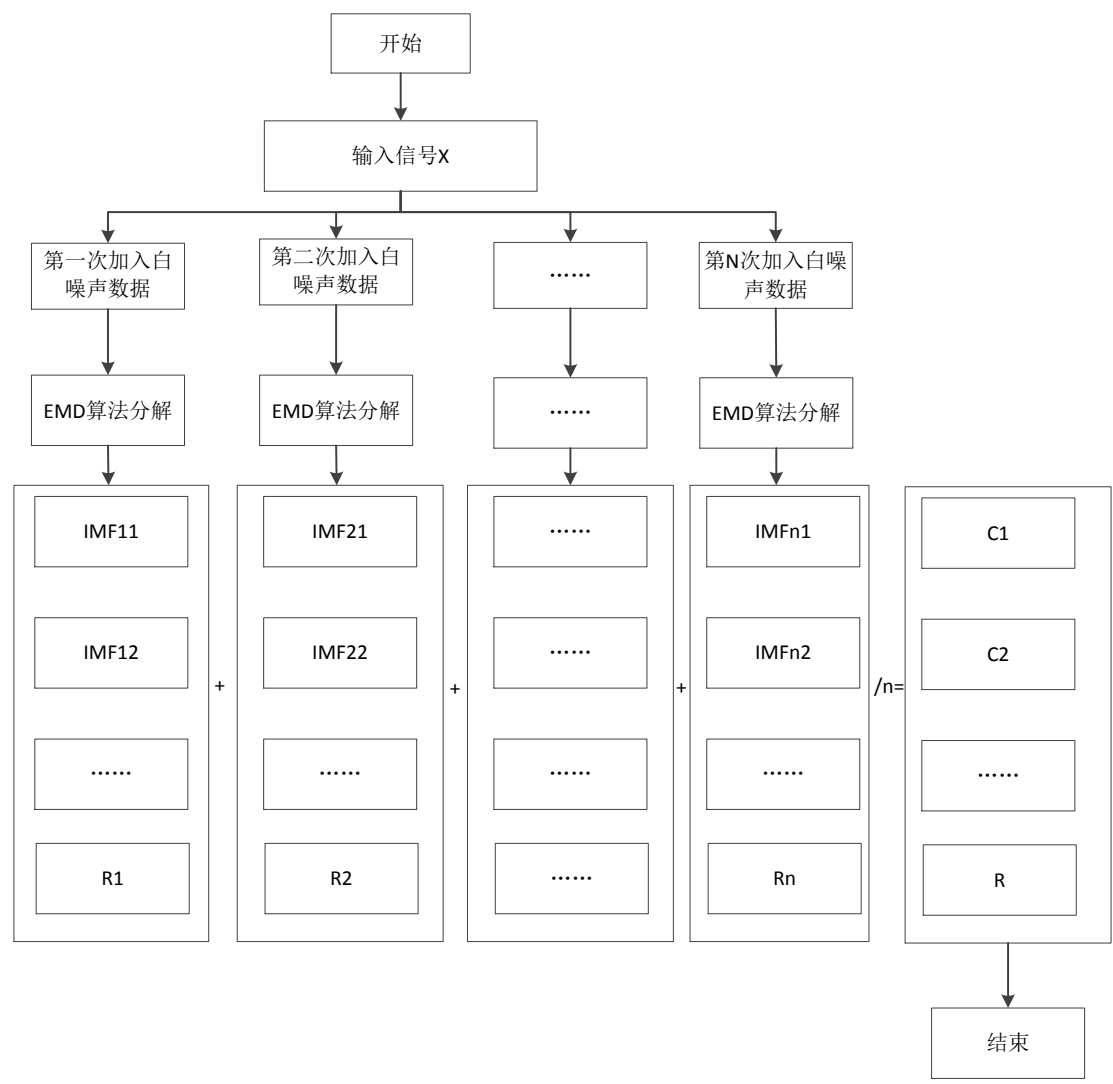


图 3. 2 EEMD 分解过程流程图

图 3.2 中， C_i 的公式如公式 3.5 所示。

$$C_i=(IMF_{1i}+IMF_{2i}+.....+IMF_{ni})/N \quad (\text{公式 3.5})$$

虽然通过添加高斯白噪声数据，EEMD 在一定程度上解决了 EMD 的端点效应问题，但是由于信号变化过快造成的模式混淆问题仍然存在，不能通过添加高斯白噪声数据解决。而电子商务环境下，短生命周期商品的销售量的随机性波动性更强，变化快速，直接使用 EEMD 算法直接分解出来的结果，可能会因为模式混淆而降低预测结果的准确性。因此不能直接使用 EEMD 算法对电子商务环境下的短生命周期商品的销量进行分解预测。

3.2 预测算法

3.2.1 ARIMR 模型

如本文第二章国外文献综述中说, Box 和 Jenkins 提出的 ARIAM 模型多在不同领域中进行时间序列数据的预测。所谓时间序列数据, 就是在时间维度上存在一定联系的数据。而时间序列数据预测, 就是根据刻画数据的历史状态, 寻找数据之间的规律性, 并建立数学模型来预测未来的走势^[41]。

对于时间序列数据有两个基本的假设:

假设一: 数据在未来的发展状态是受历史数据影响的;

假设二: 除了数据本身的因素, 数据的未来值还会受到其他的各种因素影响。

根据数据受到自身的影响而构建出的模型, 称之为自相关模型, 即 Auto Regressive Model, 简称 AR 模型。再不加限制条件的情况下, 一般认为 AR 模型的公式为:

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \dots + \alpha_p Y_{t-p} + \varepsilon_t, \text{ 记为 } AR(p), \text{ 其中 } \varepsilon_t \text{ 称为影响因子。}$$

数据受到其他各种因素的影响, 我们称之为滑动平均模型, 即 Moving Average, 简称 MA 模型。再不考虑限制条件的情况下, 一般认为 MA 模型的公式为:

$$Y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \dots + \beta_q \varepsilon_{t-q}, \text{ 记为 } MA(q)。$$

ARMA 模型称为自回归滑动平均模型, 即 Auto Regression Moving Average。它实际上是结合了 AR 模型和 MA 模型。目前, 在拟合稳定的时间序列时, 一般会择情选择这三个模型之一。

ARMA 模型的公式一般写为:

$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \dots + \alpha_p Y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \dots + \beta_q \varepsilon_{t-q}$, 记为 $ARMA(p,q)$ ^[42]。可以发现, 当 $p=0$ 时, $ARMA(p,q)$ 模型为 $ARMA(0,q)$ 模型, 相当于 $MA(q)$ 模型, 反之, 当 $q=0$ 时, 则成为 $ARMA(p,0)$, 也就是 $AR(p)$ 模型。也可以说, AR 模型和 MA 模型都是 ARMA 模型中, p, q 在特定取值下出现的状态。

所以对于稳定的数据, 需要对其进行相关性分析, 当数据受到自身影响比较大时, 也就是自相关系数较高时, 选择 AR 模型, 受到其他因素影响较大, 也就是偏自相关系数较高时, 选择 MA 模型。数据的自相关系数和偏自相关系数相当时, 选择 ARMA 模型, p, q 的取值根据数据实际情况和拟合的效果来确定。

ARMA 模型对数据的稳定性要求非常高, 而实际情况下, 多个领域的数均呈现出不稳定、离散等特征, 这使得 ARMA 模型的应用受到了限制。

于是，1970 年，Box 和 Jenkins 提出了 ARIMA 模型，ARIMA 模型的参数设置为 $ARIMA(p,d,q)$ ， d 代表差分。数据通过不同次数的差分，会去掉其中的趋势、季节等因素的影响，使得数据变得相对稳定，提高后续 ARMA 模型的预测效果^[43]。

第四章 电子商务环境下短生命周期商品销量预测方法

从第二章的算法文献综述中,可以知道,现有的研究表明,目前还没有一种完全适用于电子商务环境下短生命周期的商品的销量预测算法。总结这些算法的影响因素大致包括以下几种:噪声数据的影响,也可称为随机数据的影响;非线性数据的影响;数据不稳定状态的影响;商品生命周期短,缺少历史数据的影响。通过第三章对 EMD 算法和 EEMD 算法的介绍,可以知道, EEMD 算法改进了 EMD 算法对于数据的端点效应造成的 IMF 模型混淆问题。但仍然不能解决由于信号快速变化,给 EMD 分解算法造成的影响。

本文提出使用分段三次样条 Hermit 插值法改进 EMD 和 EEMD 算法,在生成上下包络线时,使用分段三次样条 Hermit 插值法替换 EMD 和 EEMD 算法中使用的三次样条插值法,可以有效的解决快速波动给 EMD 分解算法造成的影响,并将使用三次样条 Hermit 插值法改进的 EMD 和 EEMD 算法分别命名为 HEMD 算法和 HEEMD 算法。

本文将适用于电子商务环境下的短生命周期商品销量的组合预测算法 (Hybrid Forecasting of Short-life-cycle products in E-commerce environment),命名为 ESHF, E 代表电子商务环境,即 E-commerce; S 代表短生命周期商品,即 Short-life-cycle products, HF 代表组合预测算法,即 Hybrid Forecasting。

4.1 ESHF 算法的预测流程

通过对 EMD 和 EEMD 算法的分析提出改进的算法 HEEMD,结合使用 ARIMA 模型,最终形成了本文的组合预测算法 ESHF。ESHF 算法预测流程可以描述为如图 4.1:

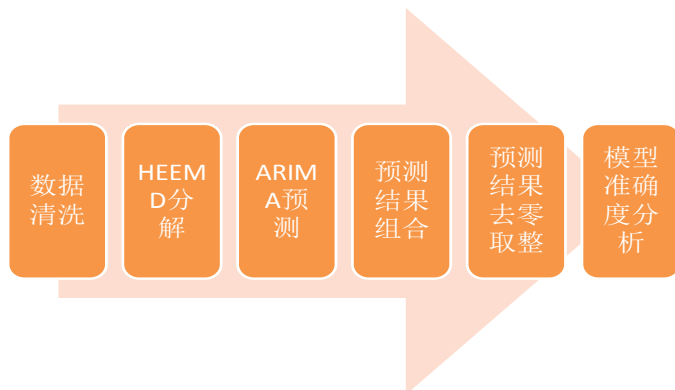


图 4.1 组合预测算法流程图

ESHF 算法分为以下几个步骤：

步骤 1：数据预处理。使用 7 点均值法取代异常数据点，降低数据的不平稳特性，为预测算法的准确度打好基础；

步骤 2：使用 HEEMD 算法进行分解。使用 HEEMD 算法进行分解的目的是，找到原始数据中的线性的稳定的 IMFs；

步骤 3：使用 ARIMA 算法对每一个 IMF 进行预测；

步骤 4：将对 IMFs 的预测结果聚合，作为原始数据的预测结果；

步骤 5：根据数据的现实意义，对预测结果进行去零和取整操作，即去除零点并对数据的小数位取整；

步骤 6：使用 MAE 和 MAPE 误差评价体系，对预测结果的准确度进行分析。

以上步骤的图示如图 4.1。

- ① 数据清洗，对离群点进行处理；
- ② 基于 EEMD，提出改进的 HEEMD 并进行数据分解，去除不稳定性；
- ③ 对每一个 IMF 使用 ARIMA 方法进行预测，并将预测结果聚合；
- ④ 后续处理，根据现实情况对预测结果去除负值并对小数进行取整操作。

4.2 数据清洗

数据清洗是对数据进行预测前的必经之路。现实数据中存在噪声点和不合理的点是非常正常的现象，尤其是像本文中所收集的电子商务环境下的销售数据，

由于生命周期短且频繁受到外部环境干扰，存在更多的噪声数据。

本文对于数据清洗，采用了与质量控制图相似的理论^[45]，假设边界条件为：

$$x_i \in [\bar{x} - 3\sigma, \bar{x} + 3\sigma] \quad (\text{公式 4.1})$$

$$\bar{x} = \frac{1}{6}(\sum_{j=3}^{j+3} x_j - x_j) \quad (\text{公式 4.2})$$

\bar{x} 是六个点的均值， σ 是左右三个点的标准差。当 x_i 不满足（公式 4.1）式时，就用

$$x = \frac{1}{7} \sum_{j=3}^{j+3} x_j \quad (\text{公式 4.3})$$

替换。从公式 4.3 中可以看出，本文所采用的替换值，是包含离群点本身在内的前后三个点，共七个点的均值。之所以没有使用前后六个点的均值替换的原因是，本文认为在商务环境下，数据的突变是体现现实意义的。比如中国电商的节日，如六一八和双十一，在这个时间段内，某些商品的销量会比较平时高出至少三倍，甚至十倍，而当过年时，商品的销量会下降到平时的十分之一左右。这种销量数据的变化是符合人们的购物习惯和现实情况的，因此应该保留数据本身的趋势特性。所以，在数据清洗过程中，使用了包含离群点本身在内的七个点的均值做替换。前文提到的不可控因素造成商品销量的暴涨或暴跌，也采用这种处理方式，保留数据本身所代表的趋势的同时，降低数据过分波动给预测结果造成的不良影响。

例如图 4.2 为商品 9783*（本文在处理商品编码时，仅显示前四位，后几位均用*代替）的异常点的修正。从图中可以看出，9783*从第十二期开始，出现了销量激增的情况，之后又慢慢回降到原始销量。根据公式 4.1，第十二个点

$$x_{12} = 11909,$$

前后六个点的均值 $\bar{x} = 5214$ ，标准差 $\sigma = 1117$ ，则

$$\bar{x} - 3\sigma = 1861, \quad \bar{x} + 3\sigma = 8568$$

$x_{12} \notin [1861, 8568]$ ，所以 x_{12} 用公式 4.3 七点均值替换，即

$$x_{12} = \frac{1}{7} \sum_{i=9}^{i=15} x_i = 6171$$

修正后的数据趋势见图 4.2。若用公式 4.2 的六点修正 $x_{12} = 5214$ ，从图 4.2 可以看

出，七点均值修正的结果相比 6 点均值修正，在一定程度上保留了数据增长的趋势。本文是要预测短生命周期商品在未来的销量情况，所以预测的结果应该是正整数。因此，对于预测结果，如果存在负数和非整数，需要对结果进行去零、去负和取整操作，取离异常数据最近的整数。

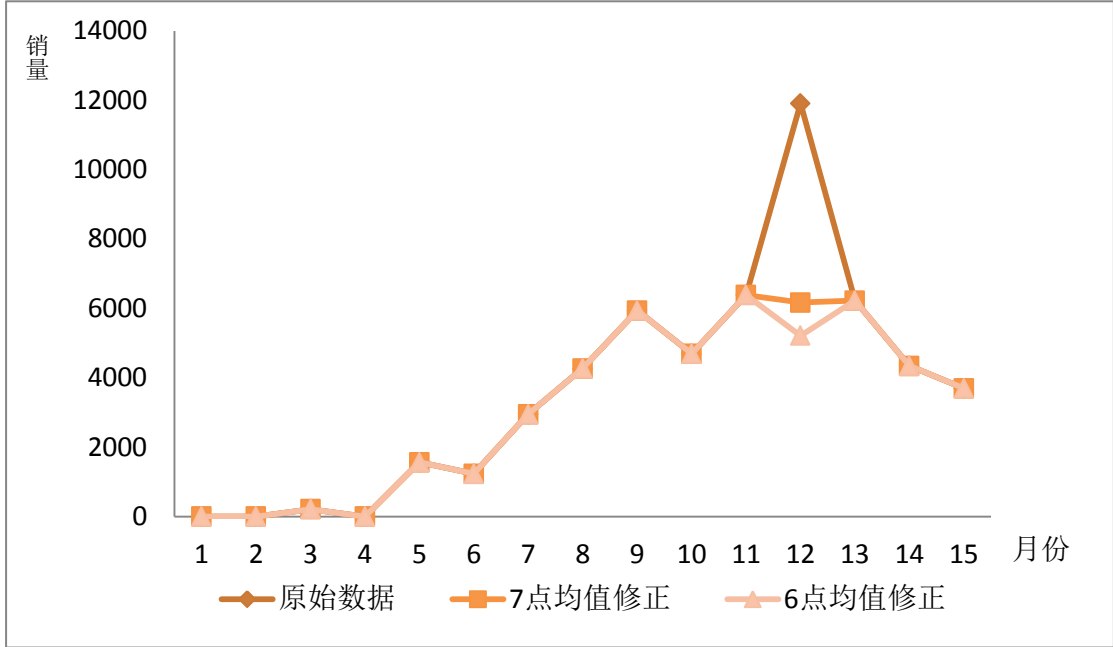


图 4. 2 商品 9375*的异常点修正

4.3 HEEMD 算法

电子商务环境下，短生命周期商品的生命周期变化快速且不稳定，因此使用标准的 EMD 算法或 EEMD 算法进行分解，得到的 IMF 均不太理想。由于 EMD 和 EEMD 均对极值点采用三次样条插值法来生成上包络线和下包络线，采用这种方法生成的上下包络线会产生过冲和欠冲现象^[46]，如图 4.3 所示，过冲和欠冲现象是指，通过极点生成上下包络线时，产生了原始数据不存在的趋势。

分段三次样条 Hermit 插值法已经被证明，可以以一定精度的逼近光滑的函数^[47]，分段三次样条 Hermit 差值法可以分段的构建出连续可导的分段插值函数，并保证生成的曲线在每个极大值和极小值之间是单调的^[48]。因此，使用分段三次样条 Hermit 插值法替换标准 EMD 和 EEMD 中的三次样条插值法，可以有效地改善生成的上下包络线的过冲和欠冲现象。

使用分段三次样条 Hermit 插值法替换 EMD 和 EEMD 中原本的三次样条插值法生成的上下包络线有效缓解了过冲和欠冲现象，见图 4.3。使用分段三次样条 Hermit 插值法替换 EMD 和 EEMD 的算法，本文称之为 HEMD 和 HEEMD。

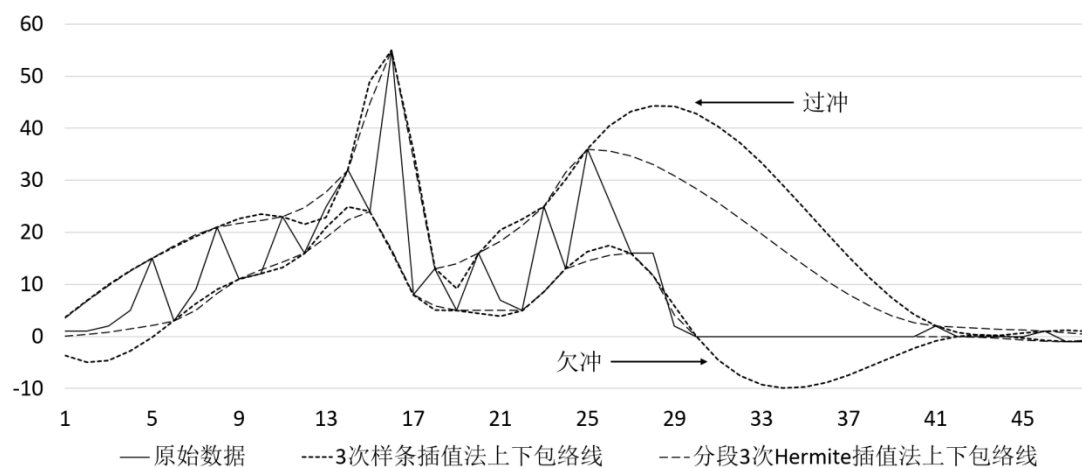


图 4. 3 由三次样条插值法和分段三次 Hermite 插值法生成的上下包络线对比图

从图 4.3 可以看出，HEEMD 算法有效的解决了 EEMD 的过冲和欠冲现象。对同一个商品真实的销量数据分别使用 EEMD 算法和 HEEMD 算法进行分解，分解结果分别如图 4.4 和图 4.5。

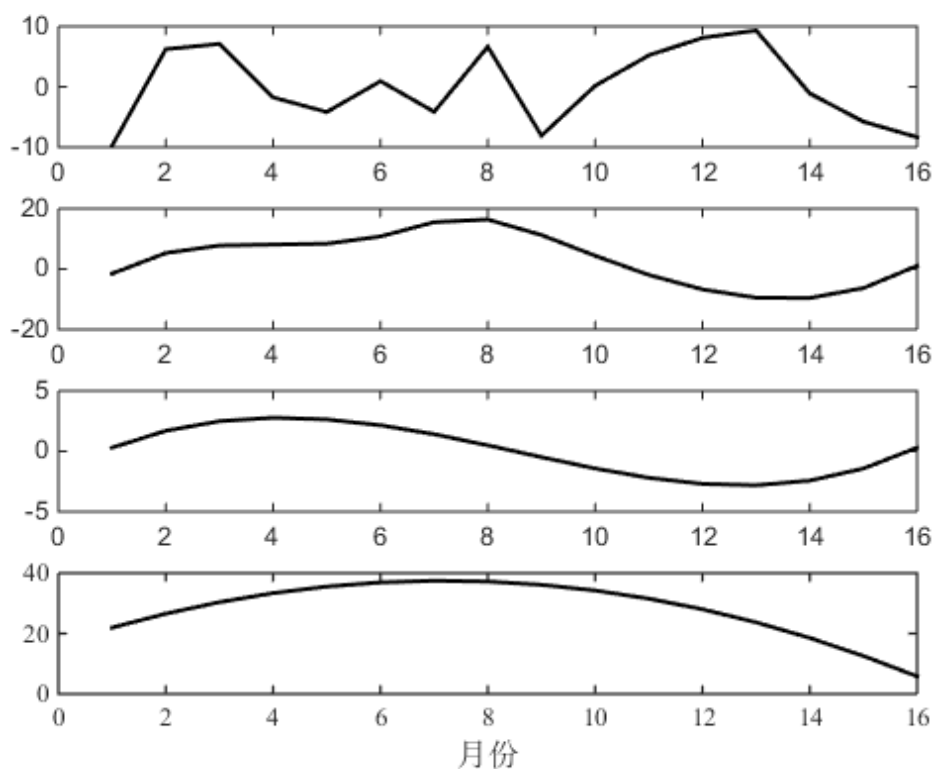


图 4. 4 经 EEMD 算法产生的 IMF 和残差

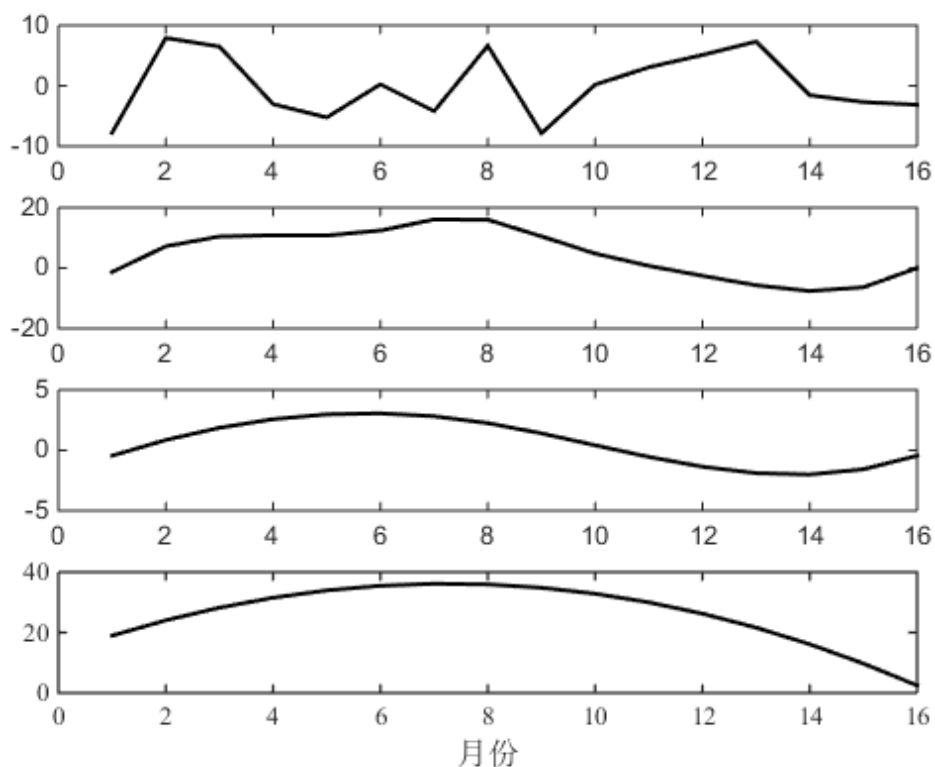


图 4.5 经 HEEMD 算法产生的 IMF 和残差

从图 4.4 和图 4.5 可以看出，数据经过 EEMD 和 HEEMD 分解后，均产生了三个 IMFs 和一个残差 R_n ，对比分解出来的 IMFs，可以看出在第 10 个月之后数据开始出现差异，经 HEEMD 算法分解出来的 IMF 更加稳定，而经 EEMD 分解出的 IMF 存在些微模式混淆的情况。现实数据的验证结果同样证明了 HEEMD 算法更适用于对电子商务环境下短生命周期商品的销量进行分解，且分解效果更好。

4.4 ESHF 算法流程

根据第三章和第四章的理论和方法检验，本文提出了基于 HEEMD 和 ARIMA 的组合算法 ESHF 算法，适用于历史数据少，波动较大的短生命周期商品的销量预测。ESHF 具体过程描述见图 4.6：

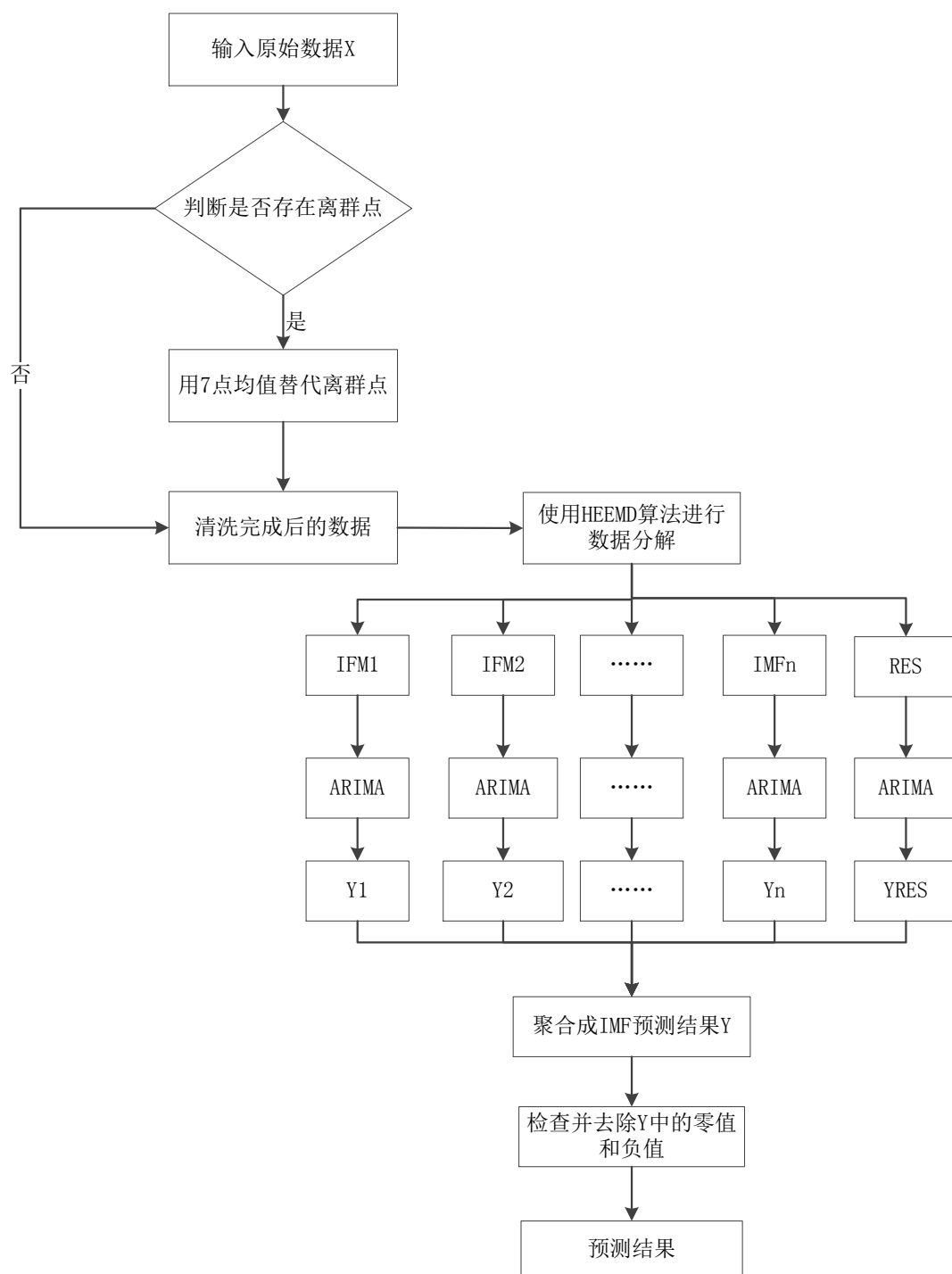


图 4. 6 ESHF 算法流程

算法使用七点均值法处理异常数据，将处理后的数据进行 HEEMD 分解，再对分解得到的 IMFs 分别使用 ARIMA 算法进行预测，组合 IMFs 的预测结果后进行去零取整操作，得到最终的预测结果。

4.5 ESHF 算法准确度的评价体系构建

为了验证组合算法 ESHF 更适用于电子商务环境下的短生命周期商品的预测, 本文将分别比对传统的预测算法 DES 算法和 ARIMA 算法的预测结果。同时为了证明 HEEMD 的分解效果比 EEMD 的分解效果更好, 本文将使用 EEMD 算法+ARIMA 算法的组合算法, 命名为 S-ESHF 算法 (Standard-ESHF, 标准 ESHF), 和 ESHF 算法的区别是 S-ESHF 算法将使用 EEMD 进行数据分解。

本文采用 MAE (Mean Absolute Error, 绝对平均值误差) 和 MEPE (MAPE : Mean Absolute Percentage Error, 平均绝对百分比误差) 两种误差评价方法来评价上述四个模型的预测结果的准确度。MAE 和 MAPE 已经在多个领域的预测中作为评价指标^[19, 49], 作为评价算法准确度的标准。MAE 和 MAPE 的值越小, 代表预测方法的准确度越高。MAE 是用来度量预测值和真实值之间的接近程度的指标, MAPE 是用百分比的形式来评价预测算法的准确程度。

MAE 的定义为:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{r}_i - r_i| \quad (\text{公式 4.4})$$

\hat{r}_i 代表预测值, r_i 代表真实值。

MAPE 的定义为:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{r}_i - r_i}{r_i} \right| * 100\% \quad (\text{公式 4.5})$$

\hat{r}_i 代表预测值, r_i 代表真实值。

当 $r_i=0$ 时, 用 r_{i+1} 替换 r_i , 则 MAPE 的公式变为:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{r}_i - r_i}{r_{i+1}} \right| * 100\% \quad (\text{公式 4.6})$$

第五章 实验设计和应用效果验证

本章通过对某电子商务公司的符合短生命周期特征的 78 组真实商品的销售数据进行预测,证明 ESHF 算法在保证商品销量预测的准确性的同时,具有可用性和推广性——可以批量解决现实的电子商务环境下的短生命周期商品的销量预测问题。本章将从解决现实问题的角度出发,使用第四章中介绍的不同的预测算法对收集的销量数据进行预测,算法是 DES、ARIMA、S-ESHF(EEMD+ARIMA 组合算法)和 ESHF(HEEMD+ARIMA 组合算法),通过两两比对预测结果的 MAE 和 MAPE,并通过成对 T 检验和 P 检验,以证明 ESHF 算法的准确性更高。

本章主要分为四个部分,第一部分介绍公司的背景和存在的问题;第二部分主要对从公司收集到的真实数据进行描述和简要分析,使用其中一个商品的销量数据作为算法应用的样例,描述算法的整个算法逻辑,最后对所有收集到的商品销量数据进行批量预测,并对预测结果进行统计学分析。第三部分从业务角度对预测结果进行分析,并对商品进行业务层面的分类,针对不同类型的商品采取不同的业务方案;第四部分为对 ESHF 算法的评价与展望。

5.1 公司简介和问题描述

5.1.1 公司介绍

J 公司是中国国内有名的 B2C 型电子商务公司。拥有自己的自建仓库和物流运输团队。J 公司以运营手机数码等 3C 商品起家,并在中国电子商务市场上,凭借优质的物流服务和正品保障,保证了市场地位。正因为 J 公司的自建仓储和物流等服务,J 公司的运营成本相对于其它电子商务公司而言更高。

J 公司在全国的仓库数量过百,实用仓库面积近两百万,覆盖全国七大区域,包括华北区、华南区、华东区、华中区、西北区、西南区和东北区,覆盖全国近两千个区县。J 公司的物流服务优质,配送迅速,在服务公司本身的同时,也向个体商家和个人用户提供快递服务。为了更好的提高用户体验,保证物流的速度,J 公司一般要求公司所经营的百万级种类的商品在全国七大仓要有现货。因此,J 公司的物流成本非常高。

对仓库进行合理配置,优化商品的摆放,可以有效的利用仓库的空间,提高仓库的使用率,但为了减低物流成本,不能仅靠仓库内优化与控制,还需要在商品进入仓库前的环节——采购环节就进行控制。以正确的数量采购“对”的商品,

将“错”的商品采取清仓手段尽快销售出去，使得仓库中的商品流动起来，也就是提高库存周转率，才能降低库存的成本。所谓“对”的商品就是能够产生销售的商品，是用户所需要的商品，而“错”的商品就是用户不需要的商品，不能产生销售最终造成滞销的商品。

J 公司一般要求库存周转天数在 30-60 天，商品的品类不同，库存周转天数会有差异。简单理解库存周转天数，就是在规定时间内，将商品卖完。这就要求采销人员能够准确预测商品在下一个周期的销量，才能够保证商品在库存充足的情况下，不至于造成滞销。J 公司经营的商品丰富，每位采销人员名下至少有 10 个品牌上百个 SKU，甚至有些采销人员名下拥有数千个 SKU。要采销人员准确的知道自己名下每一个 SKU 的库存周期天数且准确预测下一期的销量，对采销人员而言是一件非常困难且非常耗时的工作，预测商品的销量离不开运营人员的经验，了解商品和市场行情，知道商品在市场的定位，才能够准确的预测商品销量。而在人员流动过快的互联网行业内，丰富经验的采销人员数量非常少，在缺少经验的支持下去预测商品销量，极易造成预测偏差，使得商品出现断货或滞销的情况。因此 J 公司需要一个能够准确的预测商品的销量的算法，以帮助采销人员进行精准备货。

5.1.2 问题描述

J 公司运营的短生命商品的种类繁多，本文主要以短生命周期商品手机为测试品类进行销量预测。造成商品生命周期过短或历史销量数据波动过快主要有以下几个原因：

① 商品为新品。新品由于没有历史销量数据，很难预测其商品未来的特征，目前对于这类商品只能进行少量多次备货，但这样提升了多次运输的物流成本，且一旦新品成为爆品，会迅速消耗库存，造成缺货。商品缺货后，若货品补充不及时，极可能错失商品的最佳销售时期。如图 5.1 商品 9651* 的销量图，从图中我们可以看到，9651* 从第 6 个月开始瞬间下降，并持续了 3 个月，在这三个月内，商品一直处于缺货状态且供应商不能够及时到货，新货到仓后，又在 4 月内迅速销售完毕。商品 9651* 属于到货周期较长的商品，对于这类商品，如果不能一次备足货品，很容易出现断货现象。

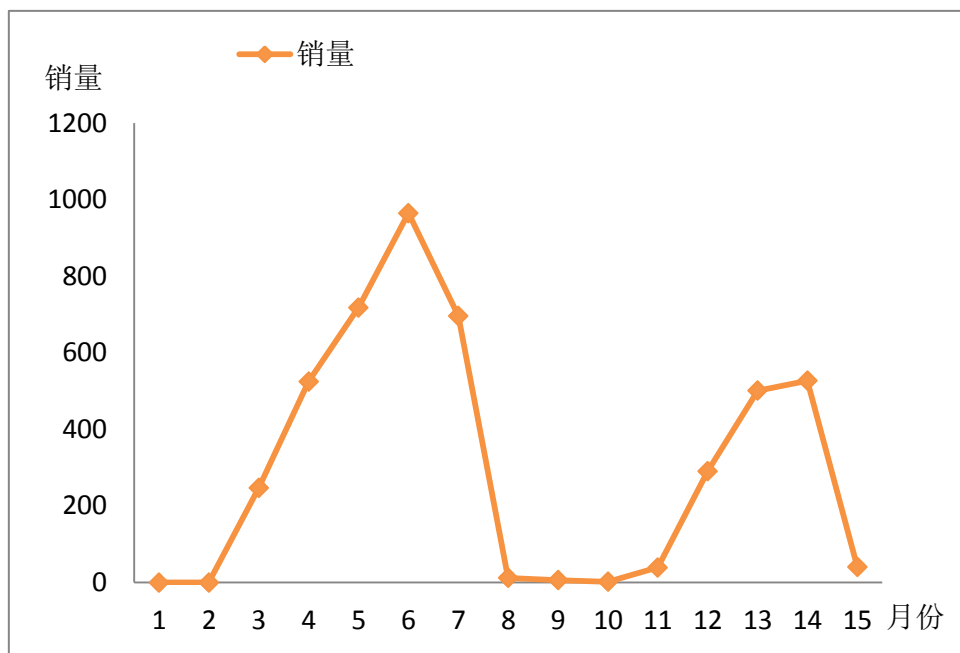


图 5. 1 商品 9651*的月销量图

② 商品历史销量波动过大的商品。从图 5.1 可以看出，商品 9651*的销量波动很大，对于这类商品，由于历史数据量较少，且波动过大，预测这类商品的销量很不容易。对于这种商品，可以看出它的整体销量趋势是下降的，中间又有 4 个月的时间处于断货期，没有历史销量，因此仅仅凭借经验和销量单维度的观察预测，会出现预测偏差过大的问题。

③ 竞品的发展和商品的升级，造成商品提前进入衰退期。图 5.2 是商品 5393*和其竞品 6562*、9783*、升级品 1057*的月销量对比图，可以看出，商品 5393*在第九个月时开始进入了衰退期，并在第十个月开始退出了市场。商品进入衰退期的原因各异，仅对 5393*而言，进入衰退期主要有两个原因，一是竞品 6562*和 9783*开始进入销售高峰期，二是 5393*的升级商品 1057*开始发力，双重原因导致 5393*迅速进入衰退期，说明竞品销量的提升和商品升级会导致商品提早进入衰退期，缩短商品的生命周期。商品进入衰退期后，就要对备货量进行控制，若不能识别商品的衰退期，并按照商品在成熟期的销量进行备货，就会出现滞销情况。商品生命周期缩短，且销量数据波动过大，都会给预测商品销量增加难度。

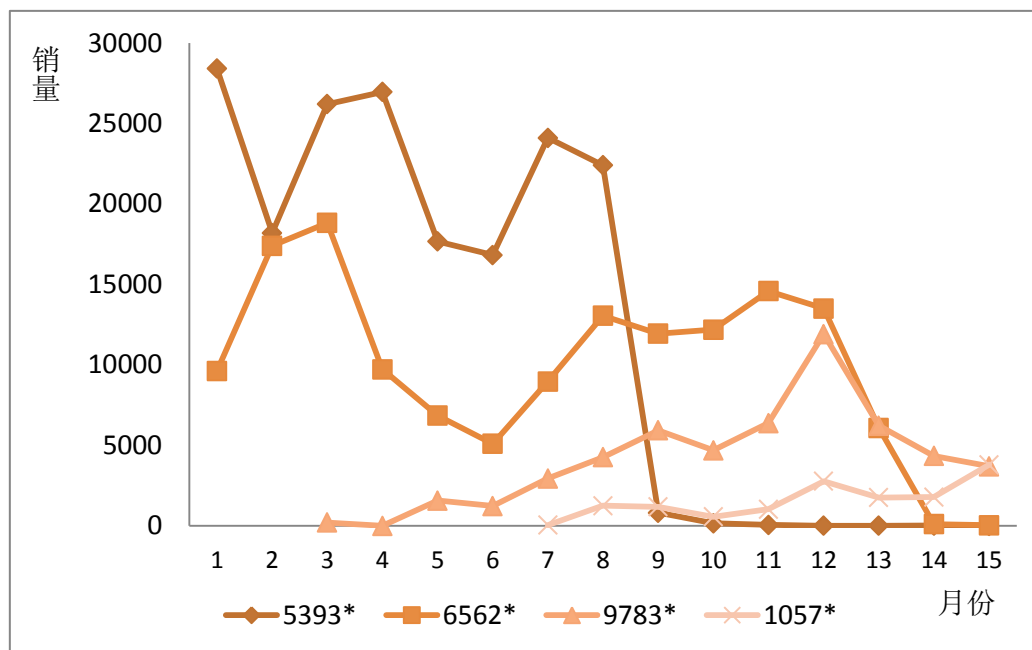


图 5. 2 商品 5393*及其互补品和升级品的月销量对比图

以上种种原因造成了商品的生命周期缩短，加大了采销人员在预测商品销量时的难度。综上所述，J 公司需要一种预测算法可以较为准确的预测商品销量，给采销人员提供可执行的备货计划。

5.2 实验设计和算法测试

本章将通过收集到的真实的商品销售数据对 ESHF 算法的准确性和有效性进行验证。主要介绍实验环境，数据预处理的方法，并抽取一个 SKU（SKU 代表一个商品，SKUID 是商品编码，具有唯一性，通过 SKUID 可以识别商品）对算法整体流程进行描述和测试，最后通过大规模的 SKU 预测，评价算法的准确性。

5.2.1 实验环境设计

处理器：Intel(R) Core(TM) i3-3110M CPU @2.40GHz 2.40 GHz

安装内存（RAM）：4.00G

实现软件：Matlab 实现算法，R studio 实现数据预处理，Office Excel 2010

进行预测结果的统计学验证（成对 T 检验）

操作系统：Windows 7 企业版操作系统

数据库：MySQL 数据库

5.2.2 数据收集和预处理

为满足 J 公司的需求，从 J 公司的数据仓库中提取某一个三级品类下 78 个

SKU 从 2013 年 7 月开始至 2014 年 10 月底，共 15 个月的销售数据。因为收集到的数据是不同 SKU 每日的销量，是一个稀疏矩阵。而 J 公司的采购期一般按月计算，因此需要对日稀疏矩阵进行按月聚合后，再进行预测分析。图 5.3 显示的从数据仓库中提取出来的原始数据，为了对原始数据进行预处理，将原始数据导入数据库中，并提取有用字段，分别是日期，SKU_ID 和销售量，并按月聚合。处理过程的 SQL 如图 5.4。

日期	一级品类ID	二级品类ID	三级品类ID	SKU_ID	订单量	销售量	订单金额
2014/10/1	9987	653	655	1023385	1	1	699
2014/10/1	9987	653	655	1084739954	5	5	881
2014/10/1	9987	653	655	1089958	34	36	10924
2014/10/1	9987	653	655	1121605	83	83	58812
2014/10/1	9987	653	655	1145230	1	1	603
2014/10/1	9987	653	655	1158184987	14	14	1088
2014/10/1	9987	653	655	1183339	1	1	240
2014/10/1	9987	653	655	1193662	5	5	8795
2014/10/1	9987	653	655	1204775	5	5	23
2014/10/1	9987	653	655	1218864	3	3	1797
2014/10/1	9987	653	655	1229091	1	1	1368
2014/10/1	9987	653	655	1251039632	1	1	78
2014/10/2	9987	653	655	1036972069	2	2	298
2014/10/2	9987	653	655	1043463	2	2	3198
2014/10/2	9987	653	655	1061621	14	14	28375
2014/10/2	9987	653	655	1128878785	12	12	2988
2014/10/2	9987	653	655	1134253	18	18	8084
2014/10/2	9987	653	655	1148568329	1	1	499
2014/10/2	9987	653	655	1148605468	2	2	698
2014/10/2	9987	653	655	1173546907	6	6	1208
2014/10/2	9987	653	655	1227193	6	6	4194
2014/10/3	9987	653	655	1075409	140	140	199560
2014/10/3	9987	653	655	1089318673	1	1	158
2014/10/3	9987	653	655	1133904155	21	21	15218

图 5. 3 部分原始数据图示

```

/*****
CREATE TABLE sku_sale_month AS (
  SELECT
    item_sku_id,
    sum(sale_qty),
    SUBSTR(sale_ord_dt, 1, 7) AS sale_month
  FROM
    origin_date
  GROUP BY
    item_sku_id,
    sale_month
)
*****/

```

图 5. 4 原始数据提取需求字段并按月聚合的 SQL

将聚合后的结果导出 Excel 后，图形化销量数据。图 5.5 是按月聚合后的结果，图 5.6 是 78 个 SKU 按月聚合后的销量趋势，从图中可以看出这 78 个 SKU 的销量并没有明显的共同特征，且波动过大，甚至有由于缺货而出现的销量断裂

的现象。

SKU_ID	销售量	月份
1217493	140184	10
1217500	102872	10
1023433	73624	10
1199748	67815	10
1057746	66323	10
1233862	51510	10
1217495	49642	10
1199988	44127	10
1139519	38972	10
1124677	37153	10
1160548	36438	10
981821	35399	10
1157477	31939	10
1222400	31161	10
1208763	31094	10
1205405	30521	10
1099152	30253	10
1150389	29859	10
1217503	27599	10
1151485	24398	10
1165981	23917	10
1217494	23612	10
1256542	22909	10
1163613	22741	10

图 5. 5 原始矩阵按月聚合后的矩阵

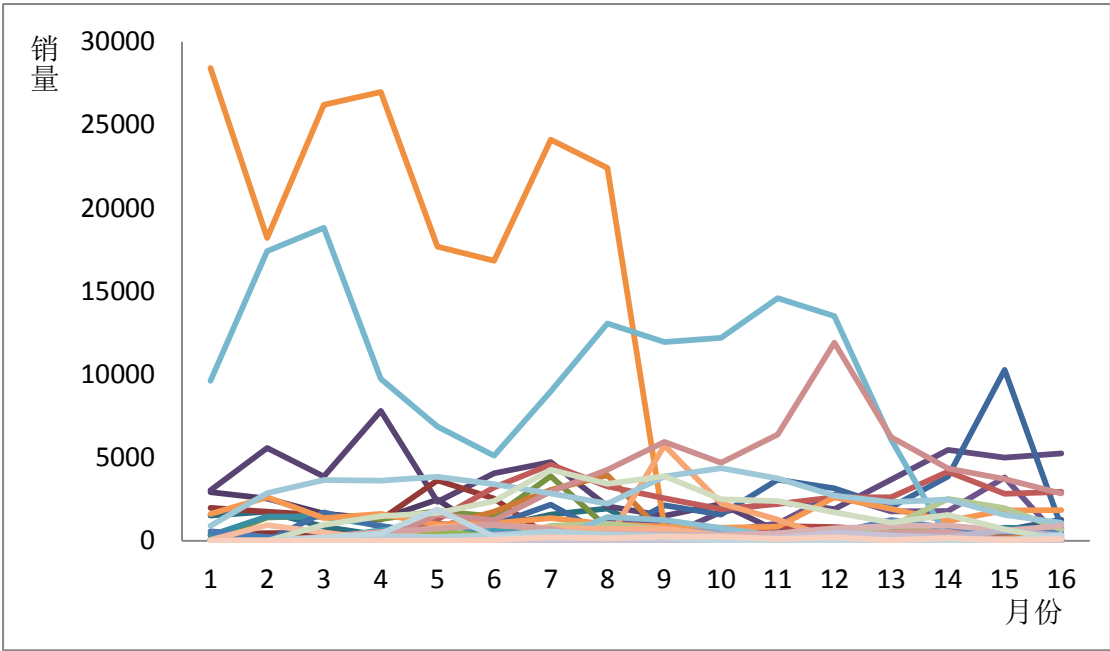


图 5. 6 78 个 SKU15 个月的销量图

为了证明 ESHF 算法可以适用于这样复杂的商品销量预测,需要对聚合后的数据进行预处理。将月销售稀疏矩阵转换成 ESHF 算法的标准矩阵,本文在数据处理过程中,采用 R 语言编写,代码见图 5.7,处理后的结果见图 5.8。

```

/*****/
data.raw <- read.csv('sku_sale_month.csv', header=TRUE, stringsAsFactors=FALSE)

sku.names <- unique(data.raw[,1])
sku.counts <- length(sku.names)
date.names <- sort(unique(data.raw[,3]))
date.counts <- length(date.names)
data.counts <- dim(data.raw)[1]

data.matrix <- matrix(rep(0, sku.counts*date.counts),
                      nrow = sku.counts,
                      ncol = date.counts)

for (i in 1:data.counts) {
  x <- which(sku.names == data.raw[i, 1])
  y <- which(date.names == data.raw[i, 3])

  data.matrix[x, y] <- data.raw[i, 2]
}

data.results <- as.data.frame(cbind(sku.names, data.matrix))
names(data.results) <- c("sku", date.names)

write.csv(data.results, 'sku_sale_month_matrix.csv')
/*****/

```

图 5.7 将矩阵转为 ESHF 标准矩阵

处理后结果如图 5.8 所示。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	sku	13-Jul	13-Aug	13-Sep	13-Oct	13-Nov	13-Dec	14-Jan	14-Feb	14-Mar	14-Apr	14-May	14-Jun	14-Jul	14-Aug	14-Sep	14-Oct
1	1000302	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
2	1000308	0	0	0	0	163	365	533	1252	131	478	723	1468	944	21	2	2
3	1000367	0	0	0	0	73	208	4	2	0	0	0	0	0	0	0	0
4	1000397	0	0	0	0	3	110	74	315	1325	1562	708	715	38	1	2	0
5	1000475	0	0	0	5	218	132	194	207	391	241	86	0	0	0	0	0
6	1000492	0	0	0	1	2607	391	1655	524	31	4	8	1305	274	404	166	7
7	1000508	0	0	0	0	3755	921	473	377	456	321	378	403	402	541	329	231
8	1000516	0	0	0	0	2293	851	537	510	448	410	253	353	328	333	222	177
9	1000530369	437	344	264	170	150	107	85	77	129	121	132	108	54	1	20	47
10	1000530566	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
11	1000530617	6	5	4	1	3	0	0	0	0	0	0	0	0	0	0	0
12	1001258	0	0	0	0	66	105	178	1966	276	79	11	11	4	8	27	1
13	1001260	0	0	0	0	48	213	89	5	8	3	0	1	0	2	0	0
14	1001262	0	0	0	0	38	71	116	412	20	3	0	2	2	0	0	0
15	1001263	0	0	0	0	139	221	320	788	660	323	185	74	36	15	1	3
16	1001271	0	0	0	0	56	155	83	22	14	4	0	2	0	0	0	0
17	1001430450	252	269	178	134	190	178	56	4	0	0	0	0	0	0	0	0
18	1001666135	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
19	1002323	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0
20	1002439355	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3	4
21	1002461717	0	0	0	0	0	0	0	0	0	0	0	3	17	18	12	3
22	1002516	0	0	0	0	1025	1374	2819	1999	1601	772	441	350	691	889	842	407
23	1002608	0	0	0	0	657	76	0	3	0	1	0	1	0	0	0	0
24	1002609	0	0	0	0	777	46	1	0	1	0	0	0	0	0	0	0

图 5.8 经处理后的 ESHF 标准矩阵

5.2.3 ESHF 算法应用实例

为了更好的展示 ESHF 算法的计算流程,本文选取了一个 SKU,商品编号 1025*作为示例,来展示算法的预测流程。SKU 1025*历史的销量情况如表 5.1

表 5. 1 1025*历史销量情况

月份	1	2	3	4	5	6	7	
销量	10	40	47	42	43	51	80	
月份	8	9	10	11	12	13	14	15
销量	61	39	38	33	27	21	5	0

首先, 根据 1025*历史销量绘制其历史销量图, X 轴为时间轴, Y 轴为销量, 可以看出, 1025*一共销售了 14 个月, 且在第 7 个月时进入销售高峰期, 或称为成熟期。但是成熟期过后马上进入了衰退期, 并在第 15 个月是时退出市场, 如图 5.9。

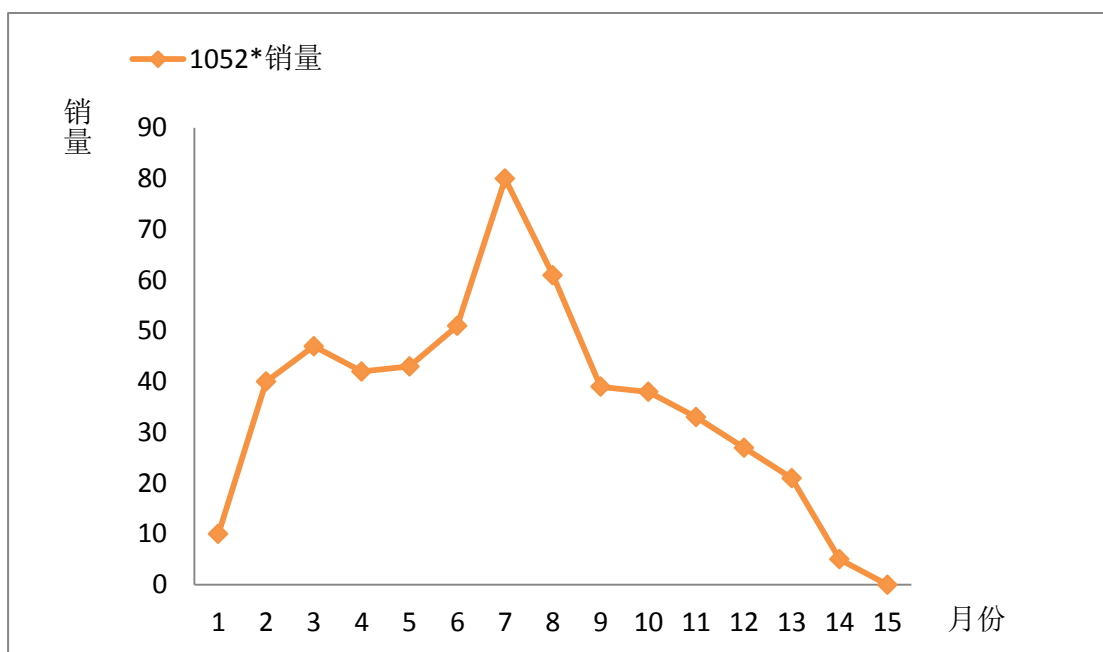


图 5. 9 1052*历史销量图

从 1052*的历史销量图可以看出, 1052*的全生命周期仅有 15 个月, 符合短生命周期的商品特征并且成熟期较短。1052*全生命周期内按月累计的销量见图 5.10。可以看出, 1052*的销量增长趋势逐渐变缓, 并在第 15 个月停止销售。

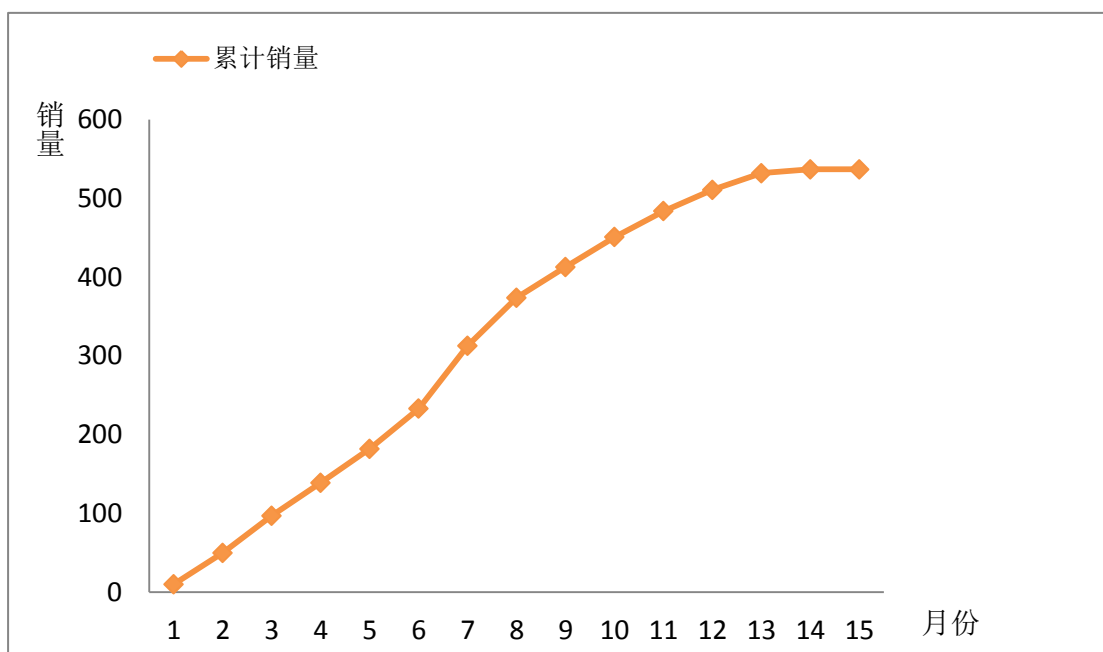


图 5. 10 1052* 15 个月内的累积销量

对 1052* 进行 ESHF 算法预测的步骤如下：

第一步，进行数据清洗。直观上看，1052*销量的离群点即为销售量最高的点，即第七月的销量。而第四到第六月份，第八到第十月份六个月的销量均值为 45.6，六个点的标准差为 8.8，根据公式 4.1 和公式 4.3 对第七月的销量进行修正，修正结果见图 5.11。修正后第七个月的销量变为 51。

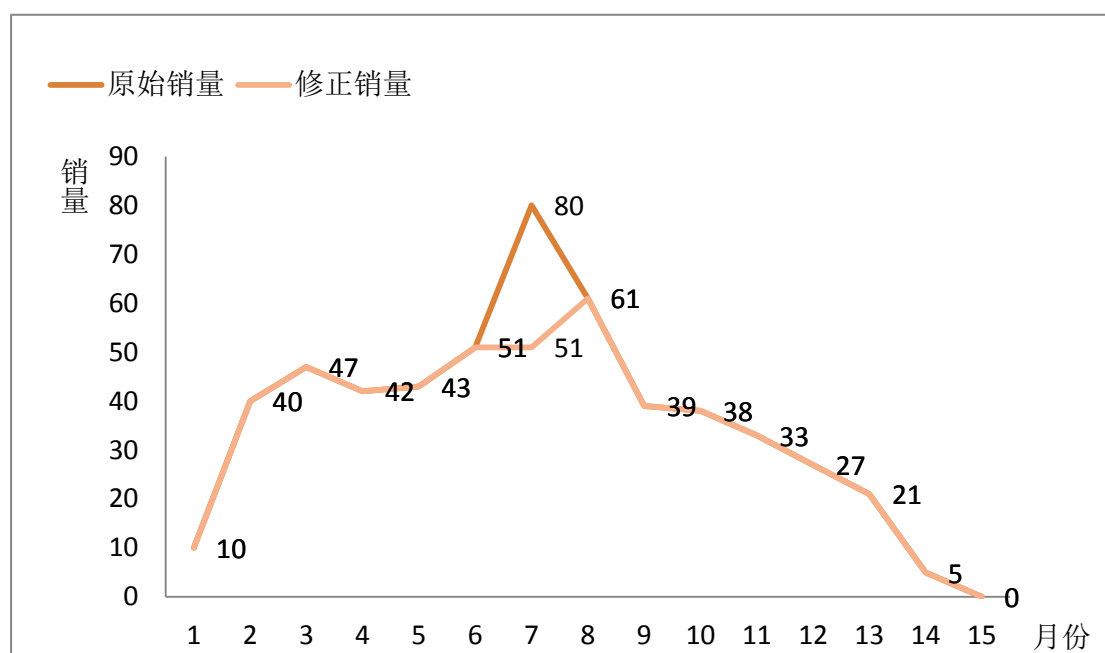


图 5. 11 1052*数据修正结果

第二步,对修正后的数据进行预测,并与原始数据进行比对。分别使用 DES、ARIMA、S-ESHF 和 ESHF 算法对数据进行预测,预测结果见图 5.12,误差见表 5.2。

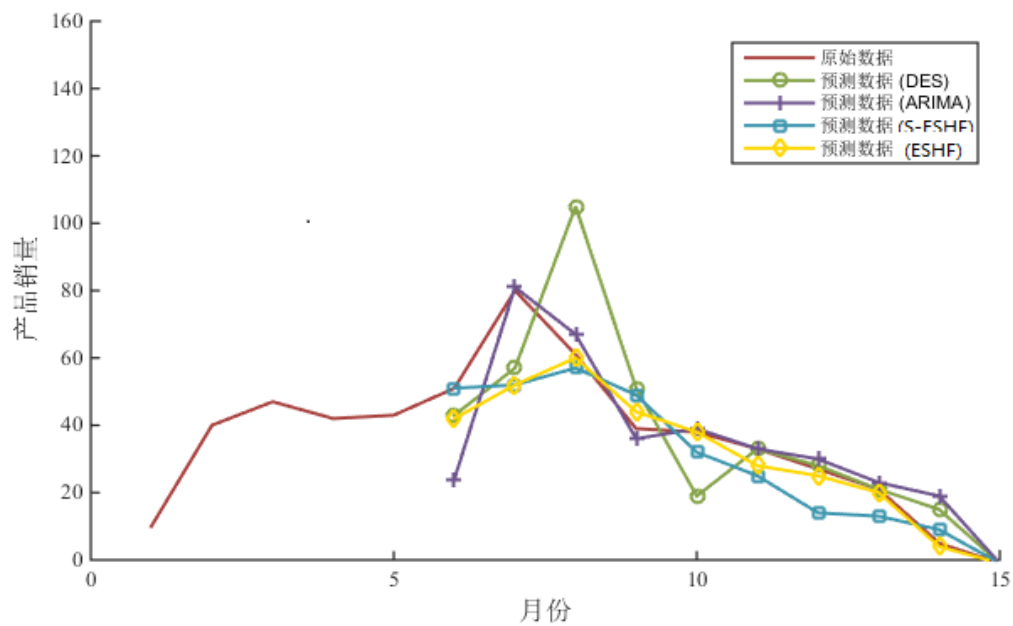


图 5. 12 四种算法对 1052*预测结果图

表 5. 2 1052*预测结果的 MAE 和 MAPE 值

	DES	ARIMA	S-ESHF	ESHF
MAE	13.000	6.3333	9.0000	5.7778
MAPE	0.4456	0.4167	0.3039	0.1271

四种算法 MAE 和 MAPE 的对比图见图 5.13。从图 5.12 可以看出算法 ESHF 的拟合效果最好,从图 5.13 中可以看出,不管是 MAE 还是 MAPE,算法 ESHF 均为最小,证明算法 ESHF 在预测电子商务环境下短生命周期商品销量时的准确性最高。

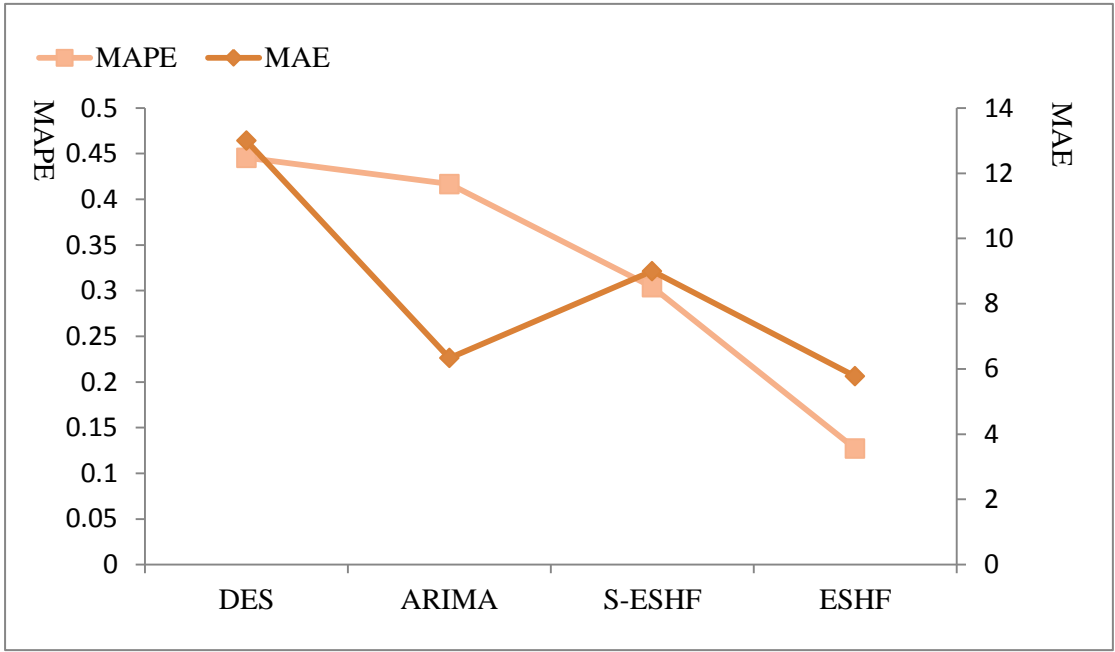


图 5. 13 1025*预测误差 MAE 和 MAPE 对比图

通过预测 SKU 1025*的销售，可以验证 ESHF 算法的可用性。为了验证 ESHF 算法的可推广性和准确性，本文将在下节采用大规模的实验数据对 ESHF 算法的准确性进行统计学验证。

5.2.4 大规模数据测试

本节将采用从 J 公司收集的 15 个月 78 种 SKU 的销量进行批量预测，并对预测结果使用 MAE 和 MAPE 进行误差评价。78 种 SKU 的 15 个月的月销量如图 5.6 所示，本文分别使用 DES，ARIMA、S-ESHF（EEMD+ARIMA）和 ESHF（HEEMD+ARIMA）共四种算法对着 78 个 SKU 进行销量预测，预测结果见表 5.3。

表 5. 3 78 个 SKU 算法的预测结果

SKU	起始月份	数据量	DES		ARIMA		S-ESHF		ESHF	
			MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE
1014*	1	16	3.89846	19.18182	1.38914	7.00000	0.68085	6.72727	1.26376	9.72727
1056*	5	6	0.64127	2.71429	0.29365	1.28571	0.39524	1.28571	0.26032	1.00000
1030*	2	1	1.46263	17.80000	0.27359	5.60000	0.20848	3.70000	0.25907	4.20000
8527*	1	5	3.15402	523.00000	65.17183	492.16667	2.60830	134.00000	6.98843	107.83333
8403*	1	3	10.72000	499.00000	29.31036	418.62500	4.37658	254.62500	5.45387	82.37500
1062*	7	6	2.54000	8.00000	0.36000	2.80000	0.94667	3.20000	0.64000	2.20000
1061*	8	7	0.64865	383.25000	0.44585	249.75000	0.20338	133.25000	0.22515	132.50000
7710*	1	5	5.85927	145.77778	42.51470	107.22222	1.68347	72.88889	6.21916	69.33333
1029*	2	8	2.17500	2.87500	0.68125	0.87500	0.25000	0.75000	0.47500	0.87500
8087*	1	0	1.22148	1200.36364	61.30512	527.72727	0.44074	262.54545	0.59292	202.81818
1023*	1	4	1.47354	4.22222	1.68704	3.44444	0.41958	2.55556	0.69444	1.88889
1065*	6	5	0.92829	201.00000	0.27068	83.50000	0.38287	97.33333	0.37810	97.00000
1064*	7	6	1.02241	18.20000	0.42708	4.80000	0.47579	5.60000	0.46672	6.00000
5678*	1	1	1.74147	443.72727	49.69808	117.54545	8.86290	36.36364	5.24498	32.63636
9277*	1	3	1.34481	840.30000	36.04180	392.00000	13.00631	307.10000	6.30195	265.60000
1089*	9	8	0.18984	963.66667	0.13858	715.00000	0.19897	1035.66667	0.18284	951.33333
1060*	6	5	0.30671	114.16667	0.30751	114.00000	0.13867	50.33333	0.04342	23.33333
1013*	6	7	2.38438	32.75000	5.55229	28.00000	0.53896	8.25000	1.83375	10.25000
1022*	6	5	3.46560	4706.16667	0.61706	1887.00000	0.26914	894.33333	0.18629	693.00000
1064*	8	7	0.87500	3.25000	1.72500	4.75000	0.87500	3.00000	0.87500	3.00000
7646*	1	5	30.39758	90.50000	16.13273	64.90000	0.42981	33.60000	0.89211	37.20000
1061*	6	5	0.36809	32.00000	0.35320	24.50000	0.18469	13.16667	0.19202	15.16667
8987*	1	6	0.75000	4.83333	17.27083	30.50000	4.40972	10.33333	5.64583	17.83333
9453*	3	7	3.71775	1221.75000	11.97151	1137.25000	12.88986	443.50000	1.34111	595.50000
9670*	3	3	0.68578	125.77778	8.94251	78.33333	2.12906	51.88889	0.32365	26.22222
1086*	9	8	0.22671	20.33333	0.19180	15.00000	0.05205	4.66667	0.09134	8.66667
1025*	1	0	0.66289	85.90909	0.32119	49.54545	0.22155	37.00000	0.13971	27.00000
1057*	7	6	26.61595	2386.80000	6.56523	1173.80000	1.59789	285.60000	1.30390	174.80000
7418*	1	5	1.00000	1.55556	0.42593	0.66667	0.70370	1.00000	0.24074	0.44444
1082*	8	8	0.29167	1.00000	0.25833	1.00000	0.21667	1.00000	0.16667	0.75000
1025*	1	8	0.30303	1.00000	0.29545	0.54545	0.05303	0.18182	0.40909	0.54545
1053*	5	4	0.37000	236.28571	0.18838	105.14286	0.34545	194.85714	0.27868	131.00000
1058*	6	5	0.43599	202.83333	0.30475	129.83333	0.12168	63.33333	0.12970	61.66667
1045*	7	7	0.90027	73.80000	13.35988	56.00000	3.01058	12.20000	0.42832	21.40000
1037*	3	2	0.68457	128.88889	0.45563	66.22222	0.22146	34.44444	0.24277	44.77778
8975*	1	0	0.39264	23.63636	0.23064	11.54545	0.10586	4.90909	0.07684	3.81818
7789*	1	8	0.78571	1.85714	0.75714	1.28571	1.41429	1.71429	3.50000	3.57143
1063*	6	5	1.31124	43.83333	1.21525	28.16667	0.59446	14.50000	0.40520	12.33333
9759*	3	4	1.37817	32.57143	3.64642	44.14286	0.83359	16.57143	0.59361	10.28571
1025*	1	2	0.44560	13.00000	0.41665	6.33333	0.30386	9.00000	0.12714	5.77778
1076*	7	7	1.45000	7.40000	0.93750	2.40000	0.53333	2.00000	0.56667	1.40000
5393*	1	0	3.56177	4058.00000	42.96617	2144.09091	2.60115	597.63636	1.07280	1054.27273
1086*	9	8	0.61762	438.00000	0.36259	232.33333	0.75101	442.00000	0.37096	229.66667
9890*	5	4	0.33407	960.85714	0.16197	471.00000	0.17002	444.28571	0.15237	377.57143
1028*	2	1	0.86139	3.20000	1.03972	2.80000	0.40194	1.20000	0.23750	0.70000
9348*	1	4	1.39698	12.71429	1.54962	6.28571	0.64738	4.71429	0.42985	5.57143
1013*	5	8	1.11905	2.42857	0.40476	1.14286	0.66667	1.42857	0.40476	0.71429
8588*	1	0	0.44943	720.00000	0.24259	319.00000	0.20382	239.27273	0.16477	218.18182
1011*	5	6	14.78333	187.40000	26.90494	210.00000	10.67870	48.80000	16.57469	104.80000
1035*	7	8	2.74400	14.00000	2.22000	7.00000	1.19800	2.00000	1.05000	2.80000
1073*	7	6	0.82910	46.80000	0.73082	44.80000	0.18541	12.00000	0.70489	37.80000
1075*	7	8	0.87279	7.00000	0.80072	7.33333	0.37802	4.00000	0.53543	5.00000
6562*	1	0	0.53295	2465.90909	14.53261	1279.90909	0.87901	803.45455	1.36790	1035.27273
1070*	9	8	0.80702	11.66667	0.84211	12.33333	6.80931	124.66667	1.00000	15.33333
1087*	8	7	1.77419	211.75000	1.03201	69.75000	0.47019	40.00000	3.05726	73.25000
9783*	3	2	0.43882	2638.33333	0.36581	2177.88889	0.10152	862.66667	0.12418	967.33333
1023*	7	6	0.89334	983.40000	0.44183	348.60000	0.25657	313.00000	0.27311	310.00000
1056*	6	5	0.55361	9.83333	0.30010	4.66667	0.29617	4.50000	0.30532	4.66667
1062*	8	8	21.72909	160.25000	16.74780	109.50000	7.50210	62.75000	9.04817	58.75000
1038*	3	2	0.67390	135.11111	0.40083	79.66667	0.31190	42.11111	0.31250	37.88889
1038*	7	6	0.90694	65.33333	0.67304	42.77778	0.20688	23.00000	0.34492	26.33333
9651*	3	2	6.84546	171.22222	1.31367	53.55556	7.74318	106.44444	3.11769	72.00000
1096*	9	8	0.66614	7.33333	0.41534	5.00000	0.38995	4.66667	0.38995	4.66667
1031*	3	8	0.87500	6.33333	1.36111	5.66667	0.75000	4.66667	0.95833	5.33333
1027*	1	0	0.26938	747.45455	0.19219	481.72727	0.10065	258.72727	0.06503	153.81818
9375*	1	4	1.59286	2.25000	0.87500	2.37500	0.87500	2.37500	8.80000	16.75000
1027*	2	5	2.65707	12.42857	2.45318	13.71429	2.25922	6.71429	1.13888	4.71429
1006*	6	5	0.35184	253.66667	0.26812	184.16667	0.16375	116.50000	0.17216	122.66667
1064*	6	5	0.90525	26.50000	0.52532	14.00000	0.40678	13.00000	0.22370	5.83333
1062*	6	6	1.03738	344.33333	0.31794	79.33333	0.42941	71.16667	0.37685	45.33333
9724*	3	4	0.39821	67.00000	0.97736	45.00000	0.31813	19.42857	0.30813	16.14286
1024*	1	3	1.50000	4.50000	1.15720	2.00000	0.61553	1.00000	1.02462	2.00000
1037*	3	6	2.44196	26.16667	2.95712	14.33333	0.87778	16.00000	0.89231	9.16667
1022*	1	6	0.95833	1.66667	0.41667	0.66667	0.08333	0.33333	0.08333	0.33333
9694*	3	2	0.53771	944.88889	1.96220	655.22222	0.81642	169.44444	0.38160	145.55556
1077*	8	8	1.00595	148.75000	2.95560	76.50000	0.48350	23.75000	1.10618	31.75000
9512*	3	6	23.50726	89.88889	33.98506	114.11111	38.93844	110.44444	15.19151	38.55556
1064*	6	5	3.35194	185.33333	0.24683	25.50000	0.38767	23.83333	0.38205	25.16667

78 个 SKU 的预测误差均值见表 5.4，图形化误差值后，见图 5.14。

表 5. 4 78 个 SKU 四种算法误差均值

SKU	起始月份	数据量	DES		ARIMA		S-ESHF		ESHF	
			MAPE	MAE	MAPE	MAE	MAPE	MAE	MAPE	MAE
均值	-	-	2.87189	397.90616	7.01989	224.79431	2.02165	122.70458	1.63877	117.47124

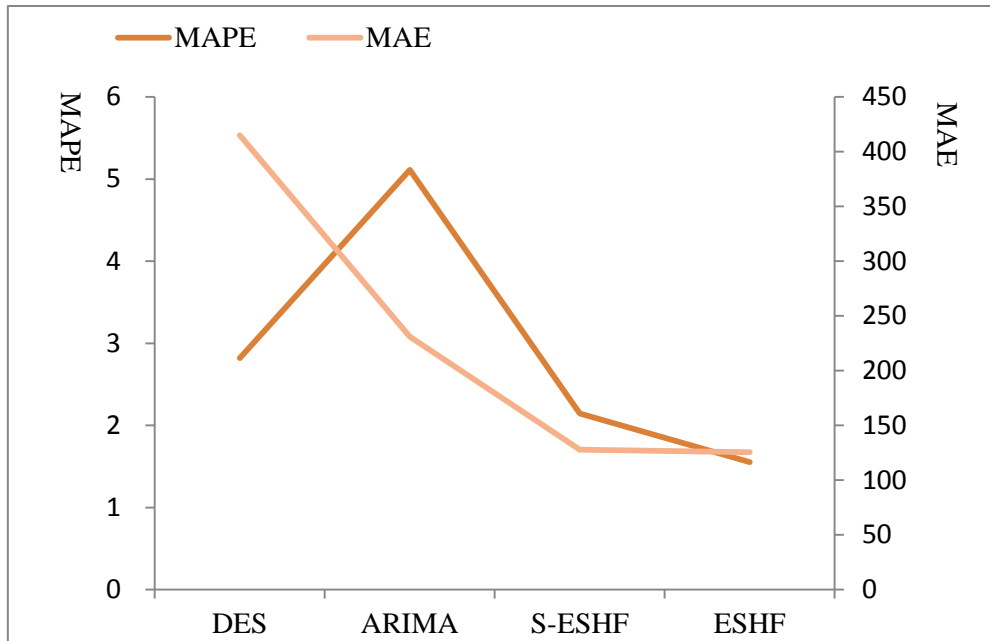


图 5. 14 图形化 78 个 SKU 的预测误差

从图 5.14 中，可以看出，ESHF 在 MAE 和 MAPE 两个评价指标下的预测误差均为最小的。所以判断在预测短生命周期商品时，ESHF 算法效果更好。为了检测 ESHF 算法在现实问题上的适用性，要对以上算法 2 的结果进行成对 T 检验。

5.2.5 四种算法预测结果分析

本节将对 DES、ARIMA、S-ESHF、ESHF 四种算法的预测结果两两之间进行成对 T 检验，本文中设置的置信区间为 0.05。假设检验的假设条件分别为：

假设条件 1: H_0^1 : $MAE = MAE_0$ ，MAE 和 MAE_0 之间没有显著差异；

假设条件 2: H_0^2 : $MAPE = MAPE_0$ ，MAPE 和 $MAPE_0$ 之间没有显著差异；

当 P 检验通过时，则拒绝原假设，认为两个算法之间有着显著的差异。

将四种算法的 MAE 和 MAPE 到入 Excel 中，准备进行统计学验证。

四种方法对 78 个 SKU 的预测结果取均值后，结果见表 5.5：

表 5. 5 四种算法对 78 个 SKU 的预测结果的均值

预测算法	MAPE	MAE
DES	2.8196914	414.96557
ARIMA	5.1134196	231.09708
S-ESHF	2.1437237	127.83608
ESHF	1.5519578	125.614

表 5.5 中,加粗的数字分别是这四种方法在 MAPE 和 MAE 两种不同评价下,预测结果最好的。可以看出,当选择 MAPE 作为预测结果的评价指标时,ESHF 算法的准确度最高,比第二好的 S-ESHF 算法准确度高出 60% 多。而采用 MAE 作为预测结果的评价指标时,ESHF 算法同样优于其他算法,因此可以确定,ESHF 算法的准确性最高。

为了进一步验证四种算法的准确程度,需对四种算法进行统计学验证。分别对 MAPE 和 MAE 进行成对 T-检验。

首先检验 MAPE。进行四种算法的 MAPE 的 T-检验验证,所使用的验证工具的 Excel,两两统计算法的预测结果,见表 5.6-表 5.11

表 5. 6 DES 和 ESHF 的 MAPE T-检验结果

	DES	ESHF
平均	2.871892201	1.63876903
方差	33.62912118	9.476587221
观测值	78	78
泊松相关系数	0.527037741	
假设平均差	0	
Df	77	
t Stat	2.209802765	
P(T<=t) 单尾	0.015044952	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.030089904	
t 双尾临界	1.991254395	

表 5. 7 ARIMA 和 ESHF 的 MAPE T-检验结果结果

	ARIMA	ESHF
平均	7.01989159	1.63876903
方差	206.7899079	9.476587221
观测值	78	78
泊松相关系数	0.553764796	
假设平均差	0	
Df	77	
t Stat	3.674957537	
P(T<=t) 单尾	0.000218847	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.000437694	
t 双尾临界	1.991254395	

表 5. 8 S-ESHF 和 ESHF 的 MAPE T-检验结果结果

	S-ESHF	ESHF
平均	2.021653537	1.63876903
方差	25.72654031	9.476587221
观测值	78	78
泊松相关系数	0.723566748	
假设平均差	0	
Df	77	
t Stat	0.952360515	
P(T<=t) 单尾	0.171947104	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.343894207	
t 双尾临界	1.991254395	

表 5. 9 DES 和 S-ESHF 的 MAPE T-检验结果结果

	DES	S-ESHF
平均	2.871892201	2.021653537
方差	33.62912118	25.72654031
观测值	78	78
泊松相关系数	0.485664805	
假设平均差	0	
Df	77	
t Stat	1.353368776	
P(T<=t) 单尾	0.089948831	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.179897662	
t 双尾临界	1.991254395	

表 5. 10 ARIMA 和 S-ESHF 的 MAPE T-检验结果结果

	ARIMA	S-ESHF
平均	7.01989159	2.021653537
方差	206.7899079	25.72654031
观测值	78	78
泊松相关系数	0.430448453	
假设平均差	0	
Df	77	
t Stat	3.388382586	
P(T<=t) 单尾	0.000555345	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.001110691	
t 双尾临界	1.991254395	

表 5. 11 DES 和 ARIMA 的 MAPE T-检验结果结果

	DES	ARIMA
平均	2.871892201	7.01989159
方差	33.62912118	206.7899079
观测值	78	78
泊松相关系数	0.304757337	
假设平均差	0	
Df	77	
t Stat	-2.660587657	
P(T<=t) 单尾	0.004744217	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.009488435	
t 双尾临界	1.991254395	

同样,对四种算法的 MAE 进行两两之间的统计学验证,结果见表 5.12-表 5.17:

表 5. 12 DES 和 ESHF 的 MAE T-检验结果结果

	DES	ESHF
平均	397.9061619	117.4712408
方差	716067.0767	57632.70834
观测值	78	78
泊松相关系数	0.836781243	
假设平均差	0	
Df	77	
t Stat	3.760752476	
P(T<=t) 单尾	0.000164165	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.00032833	
t 双尾临界	1.991254395	

表 5. 13 ARIMA 和 ESHF 的 MAE T-检验结果结果

	ARIMA	ESHF
平均	224.7943052	117.4712408
方差	208941.1297	57632.70834
观测值	78	78
泊松相关系数	0.8932255	
假设平均差	0	
Df	77	
t Stat	3.568878309	
P(T<=t) 单尾	0.000310561	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.000621123	
t 双尾临界	1.991254395	

表 5. 14 S-ESHF 和 ESHF 的 MAE T-检验结果结果

	S-ESHF	ESHF
平均	122.7045778	117.4712408
方差	48948.24097	57632.70834
观测值	78	78
泊松相关系数	0.944454439	
假设平均差	0	
Df	77	
t Stat	0.584411398	
P(T<=t) 单尾	0.280325982	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.560651964	
t 双尾临界	1.991254395	

表 5. 15 DES 和 ESHF 的 MAE T-检验结果结果

	DES	ESHF
平均	397.9061619	122.7045778
方差	716067.0767	48948.24097
观测值	78	78
泊松相关系数	0.821403498	
假设平均差	0	
Df	77	
t Stat	3.593557455	
P(T<=t) 单尾	0.000286432	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.000572863	
t 双尾临界	1.991254395	

表 5. 16 ARIMA 和 ESHF 的 MAE T-检验结果结果

	ARIMA	ESHF
平均	224.7943052	122.7045778
方差	208941.1297	48948.24097
观测值	78	78
泊松相关系数	0.853507546	
假设平均差	0	
Df	77	
t Stat	3.087876401	
P(T<=t) 单尾	0.001401377	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.002802754	
t 双尾临界	1.991254395	

表 5. 17 DES 和 ARIMA 的 MAE T-检验结果结果

	DES	ARIMA
平均	397.9061619	224.7943052
方差	716067.0767	208941.1297
观测值	78	78
泊松相关系数	0.953735355	
假设平均差	0	
Df	77	
t Stat	3.533682588	
P(T<=t) 单尾	0.00034833	
t 单尾临界	1.664884537	
P(T<=t) 双尾	0.000696659	
t 双尾临界	1.991254395	

为方便对比分析，将统计学检验结果手工整合后，见表 5.18-表 5.19 所示：
表 5.18 和表 5.19 左下角为 P 检验的结果，右上角为 T 检验的结果。

表 5. 18 MAPE T-检验结果结果整合

	均值	方差	样本数量	DES	ARIMA	S-ESHF	ESHF
DES	2.871892	33.62912	78	—	-2.66059	1.353369	2.209803
ARIMA	7.019892	206.7899	78	0.004744	—	3.388383	3.674958
S-ESHF	2.021654	25.72654	78	0.089949	0.000555	—	0.952361
ESHF	1.638769	9.476587	78	0.015045	0.000219	0.171947	—

表 5. 19 MAE T-检验结果结果整合

	均值	方差	样本数量	DES	ARIMA	S-ESHF	ESHF
DES	397.9062	716067.0767	78	–	3.533683	3.593557	3.760752
ARIMA	224.7943	208941.1297	78	0.000348	–	3.087876	3.568878
S-ESHF	122.7046	48948.24097	78	0.000286	0.001401	–	0.584411
ESHF	117.4712	57632.70834	78	0.000164	0.000311	0.28033	–

当 P 值小于 0.05 时代表统计结果显著，当 P 值小于 0.01 时，代表统计结果非常显著。从表 5.18 和表 5.19 可以看出，当选择评价指标时 MAPE 时，在 DES 与 S-ESHF、ESHF 的比较中，P 检验没有通过，也就是接受了原假设，即 S-ESHF、ESHF 算法与 DES 算法的预测准确程度无差异，也就说明改进的算法的效果并不理想；而当评价指标为 MAE 时，S-ESHF、ESHF 的 P 值远远低于 0.01，也就是拒绝了原假设，效果非常显著，并且 S-ESHF、ESHF 算法预测结果的准确度都比 DES 算法更加精确。再者，从均值上看，MAPE 和 MAE 中 S-ESHF 和 ESHF 的均值也都同时都低于 DES。所以，综合评价来看，算法 S-ESHF，ESHF 的算法准确程度要高于 DES。

比较 ARIMA 和 DES 算法的预测准确度，MAE 和 MAPE 均小于 0.01，算法结果非常显著，故 ARIMA 算法的准确度远高于 DES 算法。

比较 S-ESHF 和 ARIMA 算法的预测准确度，MAE 和 MAPE 同样通过了 P 检验，S-ESHF 的算法准确度高于 ARIMA 算法。

比较 ESHF 和 ARIMA 算法预测准确度，MAE 和 MAPE 通过 P 检验，ESHF 算法的准确度高于 ARIMA 算法。

比较 S-ESHF 和 ESHF 算法，发现均未通过 P 检验，但是 ESHF 算法均值低于 S-ESHF，证明 ESHF 算法在一定程度上，提高了 S-ESHF 算法的准确度。

综上所述，ARIMA、S-ESHF、ESHF 算法预测的准确度高于 DES 算法；S-ESHF、ESHF 算法的预测准确度高于 ARIMA 算法；ESHF 算法的预测准确度高于 S-ESHF 算法，综上所述 ESHF 算法为这四种算法在电子商务环境中，对短生命周期商品预测准确度最高的算法。

5.3 商品的分类和业务举措

虽然 ESHF 算法的销量预测结果准确度较高，但 J 公司需要准确的预测算法最根本的原因是要提高 J 公司的库存周转率，仓库的有效利用率，提高 J 公司的收益。为了实现 J 公司的目标，在完成预测后，还需要业务人员对 SKU 采取相应的措施。根据算法的预测结果和现实库存量两个指标，做成预测销量-库存四象限模型，见图 5.15，本节将介绍对处于不同的象限内的 SKU 将采取不同的业务措施。

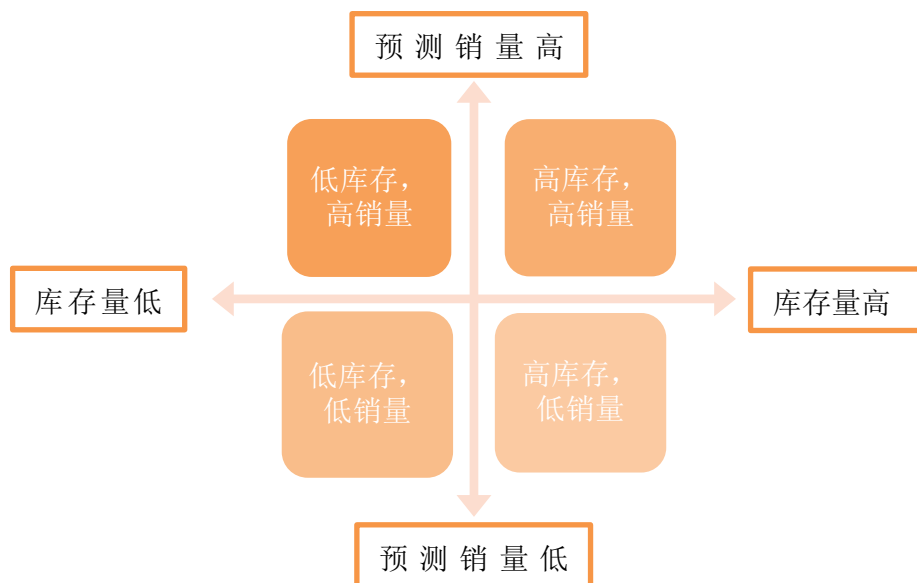


图 5. 15 预测销量-库存四象限图

对不同象限的商品采取不同措施的目的是，将商品往第一象限内引导。对于第二象限的商品，加大备货保证库存，就能进入第一象限；对于低于第三象限的商品，引导商品的销量，如销量可以上升，则会进入第二象限，此时保证商品的库存便可使商品进入第一象限，若商品销量不能上升，就让商品自然售空，防止成为滞销品；处于第四象限的商品，努力提升商品销量，若销量提升则可以进入第一象限，若销量不能提升，则容易成为滞销品。

5.3.1 预测销量-库存四象限模型应用

①高库存量，高销量

对于库存量能够满足下一期的预测销量的 SKU，且 SKU 销量的预测值不少于 SKU 的历史销量，则本文称之为高库存量，高销量型 SKU，落在预测销量-库存四象限图的第一象限。在日常运营中，对于这类 SKU 只需要保证 SKU 的正常展示，不必特意增加其流量。隔周监控 SKU 的销量，防止突然出现 SKU 销量下降或急剧上升的情况。

若出现 SKU 销量骤然下降，首先需要对标友商价格，若友商价格低于 J 公司的平台价格，则需要马上降价；其次，若价格没有问题，则需要判断 SKU 是否缺少流量，若缺少流量则需要给 SKU 提报活动，为 SKU 引来更多的流量和点击；最后，若 SKU 的流量没有问题，销量降低证明订单量下降，也就是转化率下降（转化率=订单量/流量），此时，需要判断 SKU 的页面停留时间（用户访问页面的时长），若页面停留时间下降，则需要更新商品详情页的布局，给用户带来新鲜感，带来更多的用户点击，为商品销量增加创造机会。

②低库存量，高销量

若商品的库存量低于商品的预测销量，且 SKU 的预测销量不少于 SKU 的历史销量，则本文称该类商品为低库存量，高销量型 SKU，落在预测销量-库存四象限的第二象限内。对于这类商品，需要判断两点：

第一，是否为业内爆品，所谓业内爆品，是指全平台缺货的商品。比如口罩，在雾霾高峰期时，知名品牌的口罩厂商会出现断货现象。此时，整个行业内的口罩都是供不应求的，极容易出现缺货断货的现象。对于这类商品，业务人员在提前预知会出现缺货的危机时，提早约见供应商，可以抢占先机，成功争取到更多的货源。

第二，是否为供应商管控商品。所谓供应商管控商品，是指供应商为了协调自己品牌的线上线下的销售情况，为了保证线下的销售，即使知道线上缺货也会先给线下补货。

例如某著名品牌的凉茶饮料，虽然线上的销售量很高，且长期处于缺货状态，但是厂商会首先保证线下超市便利店的库存，而不是保证线上的库存。又如某羽毛球品牌 A，为了降低线上销售对线下销售的冲击，在品牌 A 的库存销量告急时，不会先给线上补货。对于这种情况，业务人员在积极与供应商沟通的同时，要着手扶植品牌 A 的竞品品牌 B，以防止品牌 A 的断货给品牌 A 所在的类目带来商品种类的缺失或价格的断层。

如果不属于以上两点，则要尽早尽快补货，防止商品脱销。

④ 低库存量，低销量

若商品预测的销量低于商品的历史销量，且库存量不能满足商品的下一期销量，称之为低库存量，低销量型商品，落在预测销量-库存四象限的第三象限内。对于这类商品，需要判断商品的毛利是否足够高。若商品的毛利好，如果发现销量下降要及时采取类似于高销量高库存型商品类似的措施，帮助提升商品的销量。每周观察商品销量，若发现销量提升，则马上进行补货，防止缺货现象。若商品本身的毛利不高，则让商品自然售空，同时扶持毛利较高的竞品，若平台内没有类似的竞品，则需要积极引入新的竞品 SKU，来填补即将售空下架的商品的空白。弱毛利低且销量不容易提升，则直接清库，进行低价处理或作为扶植的商品的赠品，搭购销售。

④高库存量，低销量

若预测的商品销量低于历史销量，而库存量高于预测销量，则称该类型商品为高库存量，低销量型商品。对于这类商品，需要业务人员更多的关注。分析这种类型的商品，需要从以下几个方面去判断：

第一，判断历史销量好不好。若商品的历史销量本身不高，且不是新品，则

需要参加清仓促销活动，若本身进价不高，也可作为品牌赠品。对于这种商品，要采取一切措施将其销售出去，以防止停留在仓库，占用库存，最终成为滞销品。若商品历史销量好，则判断是否有新品进入平台，或竞品销量攀升。为了更好的控制库存周转，需要对比商品与竞品的库存毛利等指标，若竞品的库存不高，则限制竞品的销售，将流量引到高库存量低销售的商品上来。

第二，若商品为新品，则需要对商品进行推广。投放更多的广告，寻求更多的曝光，引来更高的流量，并时时关注商品的销量状况，若销量有明显的提升，则有望将处于第四象限的商品引入到第一象限。若商品的销量没有起色，则将商品引入第三象限后，进行清仓处理。

5.4 ESHF 算法评价与展望

ESHF 算法在准确度上，可以满足 J 公司的需要，同时结合了业务实际的操作场景，确实可以帮助 J 公司提高利润。但是 J 公司经营的商品种类非常多，超过百万级，虽然数据是 Hadoop 的底层架构，但要在 Hadoop 上实现 ESHF 这样复杂的算法对百万量级的商品进行预测，还是会使数据仓库的负载过重，有集市崩溃的风险存在。因此目前 ESHF 算法仅能在本地运行，且每次预测 SKU 的个数有上限，一般在 80 个 SKU 左右。

S-ESHF 算法和 ESHF 算法的不同之处是使用 EEMD 算法进行分解还是使用 HEEMD 算法进行分解。而在进行 EEMD 时，每次分解都会向原数据中增加白噪声数据，所以会在 EMD 算法的基础上增加常数倍的复杂度，但是准确度大大提高了。而 HEEMD 是替换了 EEMD 中的三次样条插值法，和 EEMD 算法的复杂度相同，但准确度比 EEMD 更高。为了预测算法的准确程度，牺牲常数倍的复杂度在企业中是可以接受的。

虽然 ESHF 算法存在计算复杂度的问题，但是其离线预测确实可以帮助业务人员节省工作量。结合 SKU 的库存量数据，通过 ESHF 算法可以将业务人员名下的众多 SKU 自然分层，落在预测销量-库存四象限中，并针对每个象限内不同的业务操作，实现有针对性的 SKU 的销量调整。证明了 ESHF 算法的使用价值。

日后需要进一步研究如何降低 ESHF 算法的复杂度，实现算法应用的及时性和推广性。

第六章 结论和展望

本文通过对比较短生命周期商品的生命周期与传统商品的生命周期,发现两者之间有着显著的差异。短生命周期商品除了其整体生命周期缩短之外,商品的成熟期也被压缩,且短生命周期商品一般会具有高可替代性。而通过研究电子商务公司的真实数据,发现在电子商务的环境下,短生命周期的商品容易受到网络外部因素的影响,而导致商品的历史销量不稳定,在历史销量数据少的情况下,掺杂数据不稳定的因素后,加大了商品销量预测的难度。

传统的销量预测方法一般要求商品具有较长的历史数据,且历史销量数据越稳定,预测准确度越高,而短生命周期商品的历史销量数据比较少,电子商务环境下,短生命周期商品的销量波动也很大,因此传统的销量预测方法在预测短生命周期的商品时,准确度并不理想,同样也不适用于预测电子商务环境下短生命周期商品的销量。

为了找到适用于电子商务环境下短生命周期商品预测的算法,本文比对了多种预测算法的应用条件,包括 ARIMA 和 DES,找到预测算法失效的主要原因后,针对电子商务环境下短生命周期商品的历史销量数据特征进行分解操作,使得分解后的数据适用于 ARIMA 算法,这种先分解后预测的组合算法,解决了现实环境中真实存在的历史数据量少和数据波动剧烈的问题。

本文主要通过对 EMD、EEMD 和 HEEMD 等分解算法进行研究,发现 HEEMD 的分解效果最好。因为本文提出的组合算法即为 HEEMD 分解算法和 ARIMA 预测算法的组合,命名为 ESHF 算法。为了验证 ESHF 算法的准确性和可用性,本文收集了电子商务企业中真实的短生命周期商品近十五个月的销量数据,对这些商品使用 ESHF 算法进行了销量预测,并通过统计学方法来验证 ESHF 算法的准确性。本文使用 MAE 和 MAPE 来评价算法的预测效果,实验结果证明,ESHF 算法的准确度高于传统的预测算法,即 ESHF 算法适用于电子商务环境下短生命周期商品的销量预测。

使用任何预测算法之前,都需要对数据进行预处理,尤其是对离群点的识别的处理。不同的处理方法对最终的预测结果会产生不同的影响。本文采用的离群点处理方法是七点均值法,离群点通过七点均值法处理之后,会在保留原有的趋势的前提下,降级过度波动给数据造成的影响。

为了可以将 ESHF 算法预测结果直接应用于企业员工的日常工作,帮助其节省时间同时,解决企业面临的库存问题,本文提出了业务应用模型,即预测销量

——库存量四象限模型。通过四象限模型，将 SKU 分在不同象限中，每一个象限都会有对应的业务方案，根据不同的方案，业务人员可以快速执行操作，完成合理补货，合理卖货的良性循环。

ESHF 算法实现了小规模商品的批量预测，帮助业务人员解决了部分 SKU 的库存管理问题。但由于企业经营多种品类的商品，不同品类的商品表现出的特性不同，不能直接将 ESHF 算法套用到所有的商品品类销量预测中去。为了解决这个问题，将 ESHF 算法尽量推广到适宜的品类商品预测中去，需要进一步对商品进行分类，未来希望可以实现商品的自动分类，自动预测和自动生成采购单。真正的实现自动化销量预测的同时，实现自动化合理补货。ESHF 算法已经在企业内部引起了重视，目前正在研究进一步的商品分类。相信在不久的未来，ESHF 算法一定可以实现在企业内的大规模应用。

参考文献

- [1] 焦彦军, 胡春. 基于改进EEMD方法的数字滤波器[J]. 电力自动化设备, 2011(11):64-68.
- [2] 郑称德, 王倩, 刘婉潇, 等. 电子商务市场特征对商品创新影响的实证研究[J]. 管理科学, 2014(06):90-102.
- [3] 徐学民. 电子商务中供应链管理的分析与研究[D]. 浙江大学, 2001.
- [4] 李凤. 短生命周期商品销售预测方法及其在展望电子公司的应用[D]. 河北工业大学, 2012.
- [5] 龚菲. 商品生命周期识别模型研究[D]. 南京航空航天大学, 2003.
- [6] 杨俊, 余乐安. 基于系统动力学的短生命周期商品需求预测方法比较研究[J]. 中国管理科学, 2012(S1):55-60.
- [7] 王昆强. 国际旅游岛视角下旅游电子商务的前瞻性预测分析及解决方案[J]. 改革与开放, 2009(09):66-67.
- [8] 张秋菊, 朱帮助. 基于自组织数据挖掘的电子商务客户流失预测模型[J]. 企业经济, 2011(01):95-99.
- [9] 徐琪, 刘峥. 基于SVM的短生命周期商品供应链双渠道需求预测模型[J]. 系统管理学报, 2014(02):255-262.
- [10] 刘峥, 徐琪. 基于服务-购买意愿的双渠道短生命周期商品需求预测模型——以快时尚服装行业为例[J]. 东华大学学报(自然科学版), 2013(05):668-674.
- [11] 周凯歌, 徐琪, 伍大清, 等. 周期与季节性因素影响下的短生命周期商品预测优化模型[J]. 数学的实践与认识, 2014(15):218-224.
- [12] 徐贤浩, 陈雯, 廖丽平, 等. 基于需求预测的短生命周期商品订货策略研究[J]. 管理科学学报, 2013(04):22-32.
- [13] 徐贤浩, 郭晓云. 采用Norton模型预测短生命周期商品市场需求[J]. 工业工程与管理, 2008(02):19-23.
- [14] 徐贤浩, 宋奇志. 短生命周期商品的预测方法[J]. 统计与决策, 2006(24):161-163.
- [15] Xu X, Song Q. Forecasting for products with short life cycle based on improved BASS model, 2007[C].
- [16] Xu X, Zhang H. Forecasting demand of short life cycle products by SVM, 2008[C].
- [17] Szozda N. Analogous forecasting of products with a short life cycle[J]. Decision Making in Manufacturing and Services, 2010,4(2):71-85.
- [18] Green K C, Armstrong J S. Structured analogies for forecasting[J]. International Journal of Forecasting, 2007,23(2):365-376.
- [19] Wu S D, Aytac B, Berger R T, et al. Managing short life-cycle technology products for agere systems[J]. Interfaces, 2006,36(3):234-247.
- [20] Johnson G D, Thompson H E. Optimality of myopic inventory policies for certain dependent

- demand processes[J]. Management Science, 1975,21(11):1303-1307.
- [21] Ray W D. ARIMA forecasting models in inventory control[J]. Journal of the Operational Research Society, 1982,33(6):567-574.
- [22] Miller B L. Scarf's state reduction method, flexibility, and a dependent demand inventory model[J]. Operations Research, 1986,34(1):83-90.
- [23] Box G E, Jenkins G M, Reinsel G C. Time series analysis: forecasting and control[Z]. Wiley, 200847-92.
- [24] Chung C, Niu S, Sriskandarajah C. A sales forecast model for short-life-cycle products: new releases at Blockbuster[J]. Production and Operations Management, 2012,21(5):1937-5956.
- [25] Huang N E, Long S R, Shen Z. The mechanism for frequency downshift in nonlinear wave evolution[M]//1996:59-117.
- [26] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 1998,454(1971):903-995.
- [27] Huang N E, Shen Z, Long S R. A new view of nonlinear water waves: the Hilbert spectrum[J]. Annual Review of Fluid Mechanics, 1999,31(1):417-457.
- [28] Huang N E. Introduction to the hilbert - huang transform and its related mathematical problems[Z]. 20051-26.
- [29] Rato R T, Ortigueira M D, Batista A G. On the HHT, its problems, and some solutions[J]. Mechanical Systems and Signal Processing, 2008,22(6):1374-1394.
- [30] Maa β Ss D, Spruit M, de Waal P. Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques[J]. Decision Analytics, 2014,1(1):"".
- [31] Shearer C. The CRISP-DM model: the new blueprint for data mining[J]. Journal of data warehousing, 2000,5(4):13-22.
- [32] Zhu K, Thonemann U W. An adaptive forecasting algorithm and inventory policy for products with short life cycles[J]. Naval Research Logistics, 2004,51(5):1520-6750.
- [33] Cakanyildirim M, Roundy R O. SeDFAM: semiconductor demand forecast accuracy model[J]. IIE Transactions, 2002,34(5):449-465.
- [34] 邓拥军, 王伟, 钱成春, 等. EMD方法及Hilbert变换中边界问题的处理[J]. 科学通报, 2001(03):257-263.
- [35] 刘慧婷, 张旻, 程家兴. 基于多项式拟合算法的EMD端点问题的处理[J]. 计算机工程与应用, 2004(16):84-86.
- [36] 谭善文, 秦树人, 汤宝平. Hilbert-Huang变换的滤波特性及其应用[J]. 重庆大学学报(自然科学版), 2004(02):9-12.
- [37] 赵进平. 异常事件对EMD方法的影响及其解决方法研究[J]. 青岛海洋大学学报(自然科学版), 2001(06):805-814.
- [38] Wu Z, Huang N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method[J]. Advances in adaptive data analysis, 2009,1(1):1-41.

- [39] 林近山. 基于EEMD和Hilbert变换的齿轮箱故障诊断[J]. 机械传动, 2010(05):62-64.
- [40] 张梅军, 陈灏, 曹勤, 等. 基于SVM信号延拓改进的EEMD方法[J]. 振动.测试与诊断, 2013(01):93-98.
- [41] 王瑞庆, 王晔, 李渝曾. 基于时间序列ARMAX模型的短期电价预测方法[J]. 华东电力, 2009(05):727-730.
- [42] 安潇潇. ARMA相关模型及其应用[D]. 燕山大学, 2008.
- [43] 韩超, 宋苏, 王成红. 基于ARIMA模型的短时交通流实时自适应预测[J]. 系统仿真学报, 2004(07):1530-1532.
- [44] Cooray T M J A. Applied time series: analysis and forecasting[Z]. Alpha Science International Limited, 200877-82.
- [45] Shewhart W A. Economic control of quality of manufactured product[J]. 1931,509.
- [46] Li J, Fan Y, Xu Y. An Improved Forecasting Algorithm for Spare Parts of Short Life Cycle Products Based on EMD-SVM: Information Science and Cloud Computing Companion (ISCC-C),2013, 2013[C].
- [47] 杨裕生, 盛炎平, 雷继刚. 导函数的修正三次Hermit样条插值逼近[J]. 工科数学, 1994(03):32-35.
- [48] Fritsch F N, Carlson R E. Monotone piecewise cubic interpolation[J]. SIAM Journal on Numerical Analysis, 1980,17(2):238-246.
- [49] Kurbatsky V G, Sidorov D N, Spiryaev V A, et al. Forecasting nonstationary time series based on Hilbert-Huang transform and machine learning[J]. Automation and Remote Control, 2014,75(5):922-934.

致谢

冉冉时光逝若水，匆匆岁月难回。在深夜中完成了我的这篇论文，即将搁笔的时刻，脑海里不断回想着这两年的点点滴滴，充满欢声笑语的篮球场，热闹温馨的实验室，满载着梦想和希望的自习室，和那不能遗漏的，实习期间结识的至交好友，工作期间悉心给与领导的领导。这一切的一切，让我的研究生生活幸福而充实。

在这里，首先我要感谢我的指导老师李杰教授，多年来李杰教授不仅在课业上给我指点方向，在学术研究过程中帮我突破一个个难关，让我在专业方面越来越自信；李杰教授更在生活上给予我母亲般的关怀，以身作则教导我做人的道理。是李杰教授开朗大方的生活方式，严谨认真的学术态度，教会了我认真工作，快乐生活。一路上正是李杰教授的悉心指引，才有我今天的成绩。在这里，我真心的再次感谢我的指导老师——李杰教授！

其次，我要感谢我的直属领导徐晓东经理。感谢徐晓东经理在我写论文期间，不仅帮忙把关论文质量，还帮忙协调工作事宜，给我充足的时间安心完成论文。没有徐晓东经理的理解与帮助，我就不能在兼顾工作的同时顺利的完成这篇论文。

再次，感谢我的父母，女儿多年求学在外，不能时时侍奉左右，多有遗憾，感谢多年来父亲母亲大人对我的理解与包容。

最后，感谢各位评阅本文的老师，感谢各位在繁忙的工作中抽空评阅本文，在此，我衷心的祝各位工作顺利！

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校 ☐ 一年 / ☐ 两年 / ☐ 三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日