文章编号：1005-8451（2016）09-0031-07

基于大数据平台的动态票额智能预分系统的研究与实现

汪健雄
（中国铁道科学研究院 电子计算技术研究所，北京 100081）

摘 要：本文提出了以Hadoop为核心的大数据分析平台架构，并设计了基于MapReduce的分布式预测算法、动态票额预分、敏捷票额调整等流程，构建了动态票额智能预分系统，使得客票销售形成“预测、预分、监控、调整、再预测”的闭环流程，实验数据证明，该系统可有效提升铁路客运产品收益。

关键词：大数据；12306网站；BP神经网络；并行处理；票额动态预分
中图分类号：U293.13：TP39 **文献标识码：**A

Railway Intelligent Dynamic Ticket Pre-assignment System based on big data platform

WANG Jianxiong
(Institute of Computing Technologies, China Academy of Railway Sciences, Beijing 100081, China)

Abstract: This article proposed a architecture of big data analysis platform with Hadoop as the core, designed MapReduce based distributed prediction algorithm, the process of dynamic ticket pre-assignment and agile seats adjustment, implemented the Railway Intelligent Dynamic Ticket Pre-assignment System. The System made railway ticketing form a closed loop process, which concluded prediction, pre-assignment, monitoring, adjustment, and re-prediction. The experimental data showed the System could effectively enhance the revenue of railway passenger transport products.

Key words: big data; 12306 Website; BP neural network; parallel process; dynamic ticket pre-assignment

从2011年起，铁路在全路实行旅客列车票额智能预分，采用客流预测方法生成列车席位预分方案，达到了票额管理合理化、科学化、趟车效益增加，并且自预售之日起，保证始发长途票额分配合理，兼顾沿途需求，保障中间站的旅客发送，充分提高了中间站组织客流的积极性。为各铁路局客运组织实现挖潜提效、精细化管理起到关键作用作用。随着参与预分的列车不断增多，动车组列车购票习惯的变化，现有的预分方法和实现机制也存在以下问题：

（1）铁路列车近年来调图频繁，车次急剧增加，并且预售期延长，由调图带来的停站方案、开点变更、编组调整变化较大，导致预测计算量巨大，系统负载较重。

（2）以往的票额预分为预售期外一次预测并预分，预售期内调整完全依据人工调整，不容易及时发现问题，票额调整工作被动，且临近开车期间销售情况难以掌握。

因此，有必要针对参考期内席位售出情况和预售期内余票概貌等情况进行动态监测，研究票额动态预分的方法，并对预测数据、调整依据的计算进行基础架构改造，适应海量数据变化的需要。

1 铁路客票大数据平台的研究与实现

随着客运历史数据的累积，以及全国铁路客运规模的快速扩展，全国铁路客票历史数据规模越来越大，数据种类也越来越多，仅仅依靠关系型数据库进行数据的管理和操作，已经不能满足需要。因此，以客运营销数据为基础，结合由客票生产系统产生的实时数据，采用开源分布式数据库构建大数据平台，实现铁路客票大数据平台的研究具有重要意义。

1.1 Hadoop分布式并行处理

Hadoop是近年来炙手可热的开源分布式并行处

收稿日期：2016-06-15
基金项目：国家自然科学基金(U1334207)；中国铁路总公司科技研究开发计划课题（2016X005-B）；中国铁路总公司科研计划课题（2016X004-G）；中国铁道科学研究院科研专项课题（研发中心）(J2016X005)。
作者简介：汪健雄，副研究员。

理框架，用户可忽略对底层并行实现的细节高效的构建出并行的分布式程序。Hadoop 主要包括 2 个组件：（1）与 GFS 类似的分布式文件系统，简称 HDFS；（2）并行计算模型 MapReduce，由 JobTracker、TaskTracker 等组件组成。

Hadoop 的工作原理是将数据拆成片，并将每个“分片”分配到特定的集群节点上进行分析，每个数据分片都是在独立的集群节点上进行单独处理的，因此非常适合处理大数据量、非结构化数据。Hadoop 集群的另一个特点是具有较好的可扩展性，随着数据量的增加，集群的处理能力将会受到影响，可通过添加额外的集群节点有效地扩充集群以解决问题。Hadoop 集群的并行处理能力可显著提高计算效率，能达到实时或准实时数据处理的时效性。此外，Hadoop 所需软件为开源软件，并能够很好的支持商用硬件从而客运很好的控制成本，此外，Hadoop 集群还具有故障容错的优点，当一个数据分片发送到某个节点进行计算时，该数据在集群其他节点上会保留副本，即使一个节点发生故障，该策略也能保证该节点数据的副本数据正常处理。

1.2 铁路客票大数据平台数据源

铁路客票大数据平台主要来源于历史数据和实时数据两类。历史数据包括互联网订票数据、运能数据以及售票、退票、废票和改签数据。客票系统实时数据包括实时余票数据、实时存量数据以及取票轨迹数据。其中，实时余票数据从互联网售票的余票查询集群获得，实时存量数据和取票轨迹数据从铁路局中心的客票系统获得。

客票历史数据和客票系统实时数据通过 ETL 服务，进入铁路总公司营销数据仓库，通过数据建模组成数据集市提供报表、查询应用等服务；同时上述数据也进入 Hadoop 平台的 HDFS，数据提供 Hbase 和 Hive 两种访问方式。

在票额预分应用服务层中，由客流预测应用服务器从 Hbase 中提取预测需要的样本数据，应用 MapReduce 实现客流预测算法，以实现客流预测结果。

客流预测结果通过铁路总公司客票系统服务器实现往 18 个铁路局（公司）分发。各铁路局客票系统服务器上部署预测执行子系统，将预测结果与席位实时存量数据结合生成预分方案，对铁路局中心席位库进行预分操作。

1.3 余票快照采样分析及可售能力敏捷获取

由于客票系统的分布式特性和业务复杂性，订票记录和可售能力获取目前已经达到准实时，但是为适应动车组列车的公交化开行和售票组织策略的动态调整，必须解决不影响生产系统进一步提高订票记录和可售能力获取的实时性这一技术难题。本文利用 12306 网站内存数据库技术研究了可售能力的敏捷获取方法，以实现对票额预分情况的动态调整。

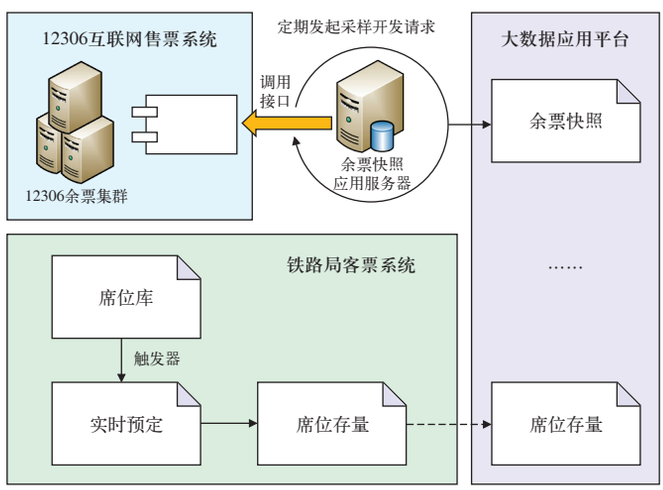

图1 余票快照采样分析及可售能力敏捷获取

如图 1 所示，在互联网售票系统余票查询集群内网应用服务中定义余票批量调用 WebService 接口，在客票网中设置逻辑上的余票快照应用服务器，需要监控的车次在监控车次定义表中做好定义，涉及到所有的购票区间、席别，应用服务器每隔一段时间向查询接口发出一次所有定义表中已定义车次的采样请求，由于车次较多，采用并发任务执行。监控车次定义表由工作流定时更新，不断新增新开的车次，取消已经停开的车次，并将车次加入预先设置好的并行分组，采样结果存入大数据分析平台。在铁路局客票系统设置席位库触发器，将订票信息实时采集到本地，结合初始预分席位生成席位存量，席位存量存入大数据分析平台，以供其他系统分析使用。

2 基于客票大数据平台的票额预分系统

基于客票大数据平台构建动态票额预分系统如图 2 所示。

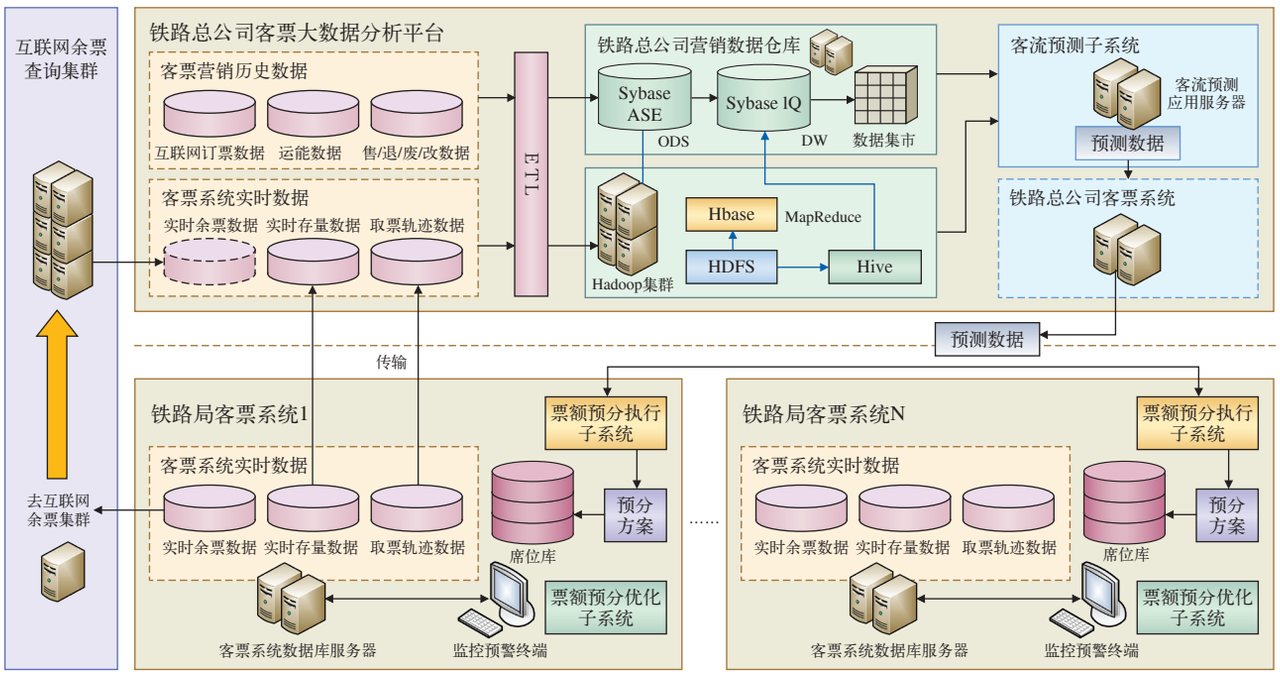

图2　基于客票大数据平台的票额预分系统

各铁路局售票历史数据通过传输软件进入铁路总公司营销系统，实时售票数据通过数据同步技术进入到铁路总公司营销系统，另外，来自于互联网售票查询集群的余票相关数据也进入到营销数据库，多个渠道的数据形成所需分析的数据源，通过 Hadoop 平台 ETL 装置进入铁路总公司营销数据仓库，在客流预测子系统中进行预测并且形成预测数据进入票额预分执行子系统，票额预分执行子系统形成预分方案通过传输下发到各铁路局形成预分方案，通过票额预分执行子系统作用于席位库，对生成的初始票额进行预分。在各铁路局通过票额预分优化子系统对预分效果进行实时反馈，形成优化方案供铁路局客运决策者进行调整，实现智能调整流程。

2.1　客流预测子系统

客流预测子系统是该系统的核心系统。历史数据是对未来计划预测的重要依据，有效数据量越大、越全面，得到的预测结果也会与实际更为接近。目前，文献中最常见的客流预测方法是外推法，该方法有很多成熟的模型，如指数平滑、ARIMA 模型、非线性回归模型、神经网络模型等。Vlahogianni, Golias and Karlaftis 指出神经网络在短期交通预测领域是最有潜力的技术，并且一些文献也归纳了神经网络的优点，如分布自由、全局最优逼近和容错性等，还有一些学者基于神经网络使用定量的方法建立了铁路客运量预测模型，因此，本系统采用神经网络构造预测模型。

2.1.1　预测数据预处理及特征变量提取

目前的历史数据都是交易数据，没有真正的需求数据，已发生的交易数据因为在实时的过程中不可避免的会产生异常值。在运用具体的预测模型时，对数据进行预处理。数据预处理主要分为 3 步：

（1）调图后对于原预测数据抽取的变化

每一次调图，都是对原有铁路列车产品的一种调整。相应的对于调图前列车的预测模型所需数据抽取，进行相应调整。运用调图历史表对各列车客流影响建模，对于进行调图后，提供预测模型所应该抽取的数据。

（2）异常值的识别、修正

数据历史交易数据在输入模型时，以月份为单位，运用分层聚类的算法，检测异常值，并用线性插值、中位数、平均数的方法将其修正。

（3）EM 算法（非限性化处理）

将上一步处理过后的数据进行非限性化处理。如果使用需求数据，这一步可省略。

通过对影响铁路客流的因素中提取特征变量，分别按日、月对列车客流进行汇总处理分别得到日

数据和月数据，发现日数据中日趋势、日周期、春暑运和小长假是 4 个主要影响变量，月份数据中，月趋势、月周期是另外 2 个影响变量，而在日数据中则不存在月趋势和月周期，春、暑运对月数据影响较大，而小长假对其影响则不显著。使用这些特征作为客流预测模型的输入变量，具体取值范围如表 1 所示。

表1　特征变量的输入集合

分析层次	特征	变量	描述	神经元数目	取值范围
日数据	日趋势	DT	反映增长趋势的连续日指标变量	1	1~365
	日周期	DC	反映每周日运量变化的周期变量	3	0，1
	春运/暑运	DS	反映春运/暑运模式的布尔变量	2	0，1
	黄金周/小长假	DG	反映黄金周/小长假模式的布尔变量	2	0，1
月数据	月趋势	MT	反映增长趋势的连续月指标变量	1	1~12
	月周期	MC	反映每年月运量变化的周期变量	4	0，1
	春运/暑运	MS	反映春运/暑运模式的布尔变量	2	0，1

2.1.2　BP神经网络预测模型

本文中，隐含层神经元的数目是输入层和输入层神经元数目的算数平均值，输出层产生预测信息并传播误差用于参数估计，输出层神经元的数目取决于预测主题的多少。根据图 3，存在 7 种类型的输入，在输入层使用了 15 个神经元，且只产生一个输出，表示为 $\hat{x}(t)$，隐含层神经元数目取 8，每个相邻层的连线具有一个权值，如图 3 所示。

该模型是基于经典 BP 神经网络模型而且是在数据输入上引入了时间特征分类，可有效地反映客流变化的各种影响因素。

2.1.3　BP 神经网络并行化

在传统数据分析平台的环境下，若对神经网络采用串行学习的方式学习效率较低，无法发挥出神经网络自身并行处理的优势，在实际应用推广方面受到很大的限制。在大数据分析情况下，传统的串行学习方式在单节点计算模式下已经无法得到满意的训练结果。随着消息传递接口（MPI）和 MapReduce 等并行编码技术的出现，可将 BP 神经网络进行并行化来提高网络的计算效率和速度。

2.1.4　BP 神经网络模型的MapReduce实现

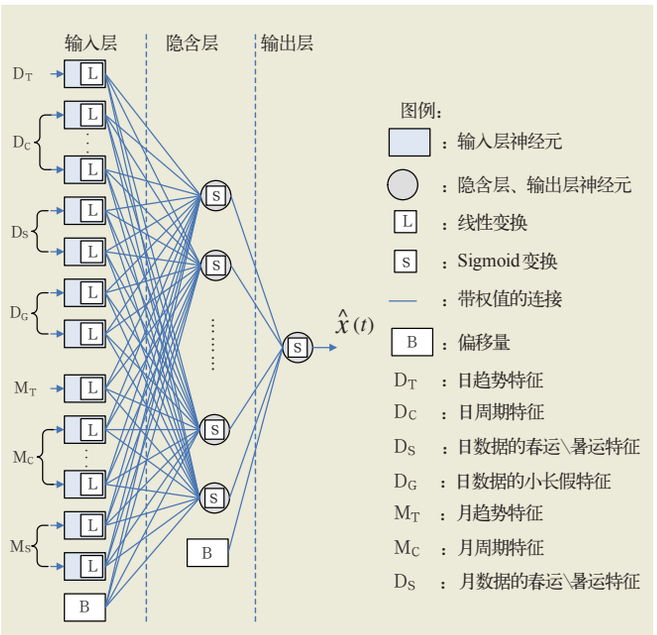

图3　BP神经网络预测模型

BP 神经网络并行化实现可通过结构并行和数据并行两种方法。结构并行根据网络结构横向分隔或者按层分隔，这种分隔方法受限于网络的节点个数以及网络的层数，当训练样本很大、网络比较复杂时，基于结构的并行化会增加处理机节点间的通信量，反而不利于计算效率的提高。神经网络并行的另一种策略是数据并行。数据并行是对样本数据进行分割，然后交由各个处理机分别进行学习训练，交换训练后的权值或者维护一个统一的权值表，可得到一个最终的网络结构。结构并行结构并不适合使用 MapReduce 编程模式实现，本文选择了基于数据并行的方法训练网络。

Hadoop 的 HDFS 会对存储的数据进行分块（Split）处理，该特点正好符合 BP 神经网络数据的特点，利用 HDFS 对数据的划分，可以将每一块数据作为一个 Map 函数的输入值，在 Map 函数内部订制 BP 神经网络的学习过程，如图 4 所示。BP 神经网络的网络的权值按一定的顺序存储在 HDFS 的共享目录中，使得每个 Map 函数都可以访问这些权值，并采用批处理方式的更新权值，每次迭代完成后对权值执行统一更新。

2.1.4.1　Map函数

Map 函数负责读取对输入的分块数据，每个 Map 函数只负责预测样本中的一部分数据。在 Map

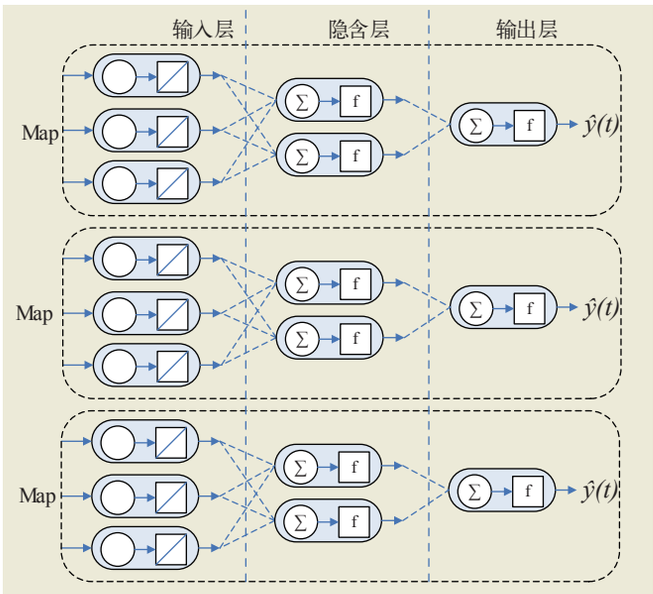

图4　并行BP神经网络预测模型

函数中提取 HDFS 中的权值，并进行神经网络正向传播，在输出层计算出实际输出值和期望值之间的差值，并根据误差计算出每个权值的改变量并作为 Map 函数的输出值。该过程的伪代码如下：

```
Input：Key-Value Pair
Output：Key，WeightWritable
i：当前迭代次数
M：本地最大迭代次数
Map(Key,Value){
    While(i<M)
     计算样本对应的权值变化量 ΔW
    }
    计算经过本地迭代后的权值总的 Σ ΔW
    创建 WeightWritable，设置权值变化量 ΔW
    Emit(key，WeightWritable)
  }
```

2.1.4.2　Reduce 函数

Reduce 函数使用 Map 函数输出的 WeightWritable 变量作为输入的 value，累计网络中的每个权值总的更新量，然后更新 HDFS 上的旧的权值，最后再将更新后的权值写入 HDFS 文件系统，在下次迭代循环开始时会使用更新后的权值进行初始化。

```
Input：Key-Value Pair
Output：更新后的权值
    Reduce(Key,Value){
        While（仍有需要处理的 Value）
            累计每个权值的更新量
        }
        更新所有权值
        将更新后的权值写入 HDFS 文件系统
    }
```

2.1.4.3　调度函数

调度函数负责启动 Hadoop 任务和控制 Mapreduce 任务的执行，并且负责初始化 BP 神经网络的权值和神经网络的最大迭代次数。在整个 BP 神经网络的训练过程中，可确保 BP 神经网络不断循环迭代。当 Hadoop 任务任务停止的情况，通常是迭代运行的次数达到了设定的阈值，或者是 BP 神经网络训练的误差降低到程序设定的可接受的范围内。

```
Input/Output：无
    i：当前迭代次数
    Mg：全局最大迭代次数
    Ev：误差阈值
    Run(){
      初始化；
      While(i< Mg)
        Map(Key,Value);
        Reduce(Key,Value);
        if（全局误差≤ Ev）break;
      }
    }
```

2.2　票额预分执行子系统

票额预分执行子系统的主要功能包括预分车次定义、预分天数定义、专家参数定义、预分方案审核、预分模板交路维护、预分方案查询及修改、预分结果查询等功能。其核心概念如下：

（1）预测数据。预测数据是通过 Hadoop 平台的 MapReduce 并行预测算法计算得出的分车次数据，其存在形式为始发站—终点站（OD）客流矩阵。

（2）预分方案。预分方案是基于预测数据生成的票额分配方案，是结合实际票额情况通过票额分配算法调整而生成的实际票额 OD 矩阵。

（3）预分模板。预分模板是历史预分方案经过专家经验确定的内置预分方案。铁路局客票管理人

员可自定义预分模板。预分模板可通过经验值人工指定，也可以通过“模板复制”功能获取一段时间内的预分数据后，参考得出模板值。预分模板分为精确模板和模糊模板，精确模板与预分方案 OD 区间一致，设置了每个预分站票额的可售区间，模糊模板是对车站分组并按以远站分块分配票额。

（4）预分方式。由于淡旺季客流的不同，决定了预分方案的不同。一般来说按模板预分管理更加严谨，而按预测预分更贴近客流实际情况。针对各铁路局淡旺季的不同，操作员可通过此功能对预分方式进行定义。操作员可以在此查询到本局所有车次的预分方式定义，并对相关车次的预分方式定义进行追加和删除，并查看相对应的操作日志。

（5）预分车次分组定义。对一些具有相同管理需求的车次，操作员可以将这些车次分成一组进行统一定义，同一组内的车次可一并添加到预分方式定义中。此功能避免可避免客运管理人员对同一类车的重复定义。

其工作流程如图 5 所示。

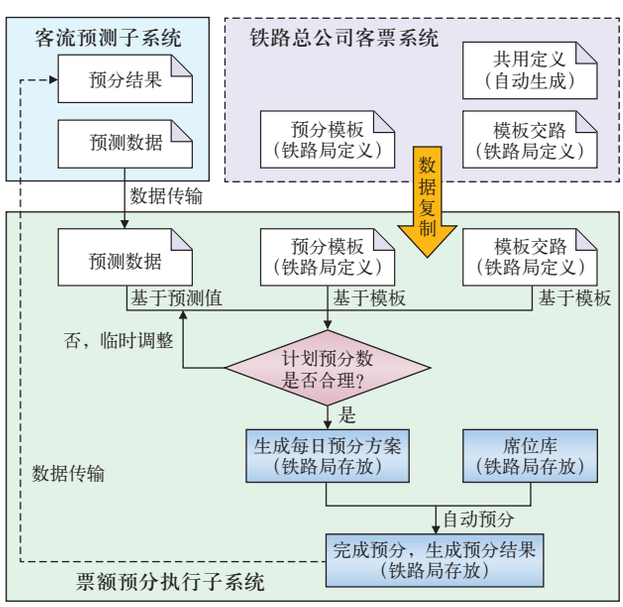

图5 票额预分执行子系统

来自客流预测子系统的预测数据，通过传输中间件到达各铁路局的客票系统生产数据库，由部署在生产库中的后台程序进行检验，判断计划预分数据是否合理，如果不合理，则需要进行临时调整，如果判断预分数据合理，则生成每日预分方案，将预分号的列车席位存放在铁路局生成数据库，同时将预分结果记录在预分结果表中，再回传至票额预分优化子系统。计划预分的数据也可以来源于铁路局客票生产库中的预分模板和模板交路，这样可以得到一个相对稳定的预分方案。

2.3 票额预分优化子系统

2.3.1 动态票额预分

由于客票系统预售期较长，传统的票额预分方案是基于预售期外 1 次预测结果生成的，预售期之内不再重新预分，因此，无法适应预售期内偶然事件的影响。从 2014 年开始，票额预分系统引入了动态票额预分，可在预售期内进行周期性的动态客流预测及多次动态调整，如图 6 所示。以 2014 年 6 月 17 日为例，这一天预测子系统将产生 2014 年 7 月 10 日始发列车的 OD 客流预测，同时调整 2014 年 6 月 30 日和 2014 年 6 月 23 日的始发终到预测数据（这两日初始预测数据分别在 2014 年 6 月 8 日和 2014 年 6 月 1 日生成），在票额预分执行子系统中将预分 2014 年 7 月 6 日始发列车的席位，并对 2014 年 6 月 29 日和 2014 年 6 月 22 日始发列车的票额进行重新预分。

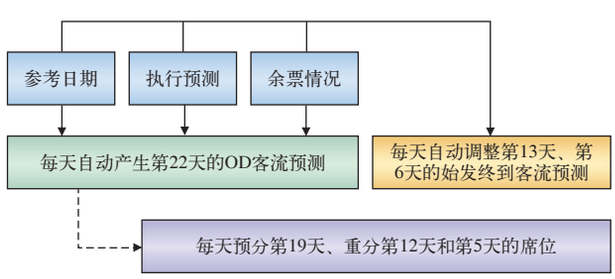

图6 动态票额预分

票额动态预分是基于客流按周变化的规律较为显著的特点进行的。在预售期为 20 天时，最多通过 3 次预分即可达到非常满意效果，但在预售期延长至 60 天的时候，由于客流变化较大，且高铁、城际列车在开车前一日和当天的预售情况变化非常显著，仅靠预售期之外的动态调整也不能很好的满足预测需求，结合余票快照分析技术实现敏捷票额调整。

2.3.2 敏捷票额调整

余票快照分析模块能记录每个时刻余票历史截面的可售能力。由余票快照分析模块取得的余票情况可通过图表观察得知，图表的横坐标为观察日（观察点），纵坐标为对应的观察点的余票快照数据。一

条折线表示对应某一下车站的余票变化趋势。余票波动图用于显示在车次、日期、席别、上车站确定的情况下，到各站的可售剩余票数随时间的变化情况。在预售期内距离发车时间 3 天以外的取数时间间隔为 1 天，3 天以内的时间间隔为 1 h。

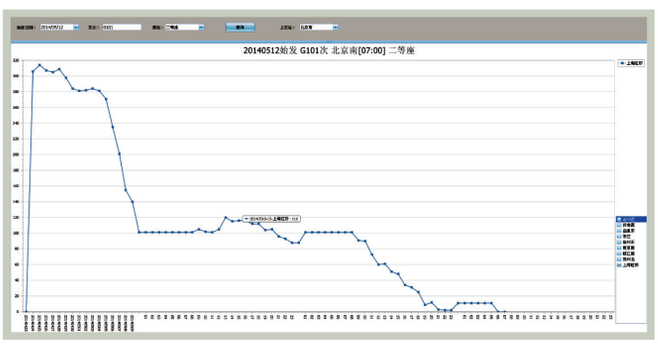

图7　票额预分余票快照波动图

图 7 为 2014 年 5 月 12 日 7:00 始发的 G101 次列车各区间的余票消逝情况，图 7 中默认为北京南—上海虹桥这一始发终到区间的余票，可得知该区间首次售完在 2014 年 5 月 11 日 23:00。说明次日首列始发的京沪高铁动车始发长途票在前一日晚间 23:00 全部售罄，由于首班高铁旅客一般不会在开车前即买即走，而夜间高铁旅客购票相对较少，相当于既能保证始发长途票在开车前有票可买，又能保证始发长途票及时卖完。因此该结果符合预分的初衷。若开车前始发长途票既未卖完，而沿途区间在开车前一直无票可售，则说明始发长途预留过多，因调配一些到沿途站销售。

3　结束语

实际应用中 Hadoop 集群使用了 16 台 HP DL380 的服务器，操作系统是 RedHat 6.4，每台服务器上安装了 JDK1.6 和 Intel 的 Hadoop 稳定版 IDH2.3。16 台服务器中，1 台机器作为 Master 节点，剩余机器作为 Slave 节点。客流预测子系统开发环境采用 Eclipse，开发语言使用 Java；票额预分执行子系统前台应用采用 PowerBuilder 开发，与客票核心系统保持一致；预分优化子系统采用 .net 开发。

通过对京沪、京广等干线经过一段时间的试用及跟踪分析，可看出旅客发送量、客运收入都有 5% 以上的提升。尤其是在传统的客运淡季，其增收的效果更为明显，如图 8、图 9 所示。

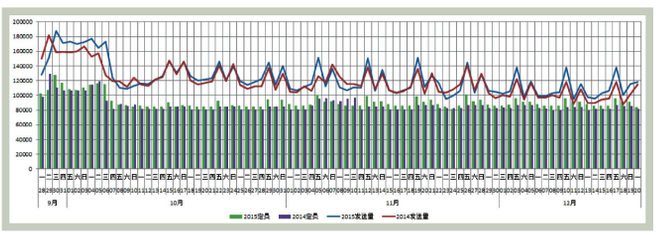

图8　京沪高铁本线列车发送量对比

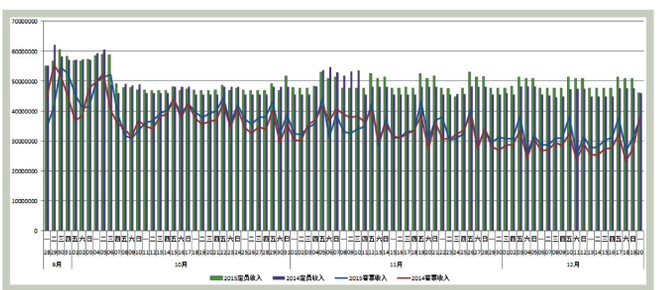

图9　京沪高铁本线列车收入对比

在铁路运输企业改革推动下，铁路客运业务快速发展，对新一代客票系统对票额管理精细化和智能化以及提高铁路运输企业效益等方面提出了更高的要求，基于大数据平台构建了动态票额智能预分系统，形成了“预测、预分、监控、调整、再预测”的闭环流程。进一步提高了票额预分系统的可用性和有效性，为铁路实施收益管理提供理论依据和技术储备。

参考文献：

[1] 胡志鹏，王洪业，王元媛，等．铁路客票席位智能预分的设计与实现 [J]. 铁路计算机应用，2016，25（1）：30-33.

[2] 王元媛，单杏花，王洪业，等．铁路客票票额预分管理系统的设计与实现 [J]. 铁路计算机应用，2015，24（12）：22-27.

[3] 单杏花，周亮瑾，吕晓艳，等．路旅客列车票额智能预分研究 [J]. 中国铁道科学，2011，32（6）：125-128.

[4] 汪健雄，贾新茹，王炜炜，等．铁路综合客流与列流管理系统的研究与实现 [J]. 铁路计算机应用，2014，23（12）：23-27.

[5] 汪健雄，张军锋，王炜炜，等．一种改进的神经网络预测模型在铁路春运客流预测中的应用 [J]. 中南大学学报：自然科学版，2011，42（S1）：1020-1025.

责任编辑　徐侃春