DOI:10.16836/j.cnki.jcuit.2015.05.007

文章编号: 1671-1742(2015)05-0445-05

基于 Spark 计算模型的随机森林的电话量预测研究

王 琪, 张洪伟
(成都信息工程大学计算机学院,四川 成都 610225)

摘要: 为提高电话中心通话量的预测效果,提高模型预测的召回率、精度、F 值,使用基于多个决策树的随机森林模型和 Bagging 方式组合,通过多个弱分类器,最后汇总。随机森林在 Spark 数据引擎中易于并行化。Spark 的特点比 MapReduce 数据引擎更适合做迭代运算和交互式挖掘。将平安科技电话中心通话量的预测精度提高了5 %。研究结果表明,基于 Bagging 组合的随机森林算法会提升数值型模型预测的效果。

关 键 词: 计算机应用; 智能工程; Spark; Hadoop; 通话量预测; 随机森林
中图分类号: TP311.13 文献标志码: A

随着通信的发展,各种业务的发展,各个中心的话务人员越来越多,按传统的经验进行排班,会导致人力资源的浪费或紧缺。电话量是安排话务员资源的前提,各大呼入中心可以针对不同的通话量安排对应数量的话务人员,在服务水平不降低情况下,优化资源配置,合理规划企业结构。因此需要准确的预测通话量。当前的 Spark 计算模型在计算,迭代计算的数据挖掘场景下表现优异。以后通话量场景的数据会越来越多,非结构化数据会越来越复杂,所以使用 Spark,对于以后的扩展很方便。机器学习算法大多数在内部实现的时候,都需要进行大量的迭代运算,所以 Spark 特性,使 Spark 特别适用大量迭代运算的机器学习算法。

1 Spark 介绍

1.1 Spark 简介

Spark 是针对大规模数据处理的一种快速和通用的数据引擎[1]。有基于内存计算的特性,可以减少 IO 操作,大大提高数据处理的速度。以前基于 Hadoop 的大数据处理架构,针对各种场景,需要使用多套系统,MapReduce 适合于做批量工作,Storm 应用于流式计算,Mahout 做数据挖掘,hive 做数据仓库。这样,在多个应用系统之间,必须要进行数据的存储格式切换,这样的模式会增加系统的复杂程度和压力。但 Spark 基于 Resilient Distributed Dataset,能够适用于各种应用场景的混合计算。且 RDD 采用了记录操作日志来实现容错,避免 checkpoint 方式的网络和磁盘开销。这些特性允许 Spark 在一个程序中,应用迭代计算、图型的计算、流式计算等,方式更灵活多变。

1.2 Spark 的各个组件

使用 Scala 编写算法。因为在解析 Scala 和实现算法的时候,用 Scala 速度最快,切合度最好。Spark 的生态圈种类繁多[2],如图 1 所示。

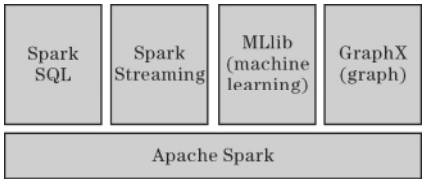

图 1 Spark 的生态系统图

有用于处理结构化数据的 Spark SQL,类似与 Hadoop 的 hive,把不同数据源数据转化成统一 SchemalRDD,这样能够使用统一的格式,高效率访问和处理不同的数据源。

有用于数据挖掘的 MLlib,在 MLlib 库中包含了许多常见的机器学习算法,常见算法包括经典的罗辑回归、决策树(是 C4.5 算法)、贝叶斯等。

GraphX,可以做图形计算,转换成 RDD,再做一些操作,实现图算法。

还有做实时流式数据的 Spark Streaming,其原理是把流式数据转化成一小块 RDD,每次只处理这小块的数据集。

Spark 对开发人员提供了 3 种语言: 支持最好的原生 Scala,使用广泛的 Java 和最近比较火的 Python。这里使用 Scala,兼容性更好。

1.2 Spark 的 RDD(弹性分布式数据集)

Spark 提出一个高度抽象的概念 RDD,相当于一

收稿日期: 2015-09-28
基金项目: 国家自然科学基金资助项目(61502086)

个记录集合[3]。RDD 也可以如 HDFS 在稳定存储中读取数据而创建。Spark 把要处理的数据，处理的中间结果和输出都定义成 RDD。且为处理大量的数据，还把 RDD 内的数据进行分区，分散到多台节点上，以便之后并行处理数据。RDD 默认是存在内存中，只有当数据大于 Spark 设置的阈值时，才会把数据溢写到磁盘。Transformation 可以看作是一组数据处理到另一个阶段的状态，如 flatmap，groupby 等操作，并没有实际产生数据，Action 是前面多个状态到另一状态的操作，如 collect，count 等操作。大部分业务逻辑或者数据处理都可以用 DAG 表示，有不错的容错机制。Transformation 也叫做 Narrow Transformation，窄依赖，Action 也叫做 Wide Transformation，宽依赖。图 2 是 Spark 的调度情况。

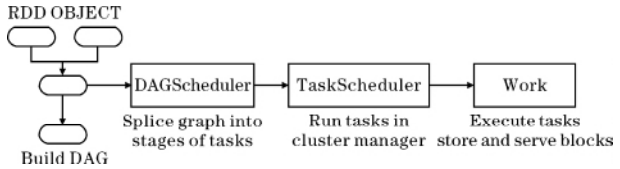

图2 Spark job 的调度

2 通话量预测概述

2.1 通话量概念

通话量是用户接入到呼叫中心的次数，反映用户与中心的交互。在网络发展的背景下，人们越来越愿意通过电话处理自己对应在该公司的服务。公司提供这种人性化的安排，大大满足用户对业务的需求，提升用户在本场景中的体验，但是也给自己带来了人员配置、网络流量等方面的压力。如果话务员安排过多，会导致资源浪费，如果安排过少会导致用户体验下降，甚至流失这部分用户。所以预测通话量，能优化资源配置，在最大限度满足用户需求的同时，实现资源的合理利用。

2.2 常见的预测方法

预测就是对还没有发生的事物进行预先的估计和判断，是通过以前发生的事物和现在发生的事物对以后发生的事物的一种探索。根据不同的需求和场景的复杂性，预测分为分析预测法和技术预测法[4]。

分析预测法是制作一个与相关事件的数学模型

$$T_{M+m} = T_m + C\sum_{u=m+1}^{M}\Delta G_u \quad (1)$$

假设 T_m 代表的是第 m 天的通话量，预测第 M 天的通话量，C 是一个常数，ΔG_u 是从 $m+1$ 天到第 M 天的通话量增长，还有其他因素，人口统计，工业，收费等。国内外有许多场景已经成功使用分析法，取得比较好的成果。分析模型是经营战略。

技术预测法是把预测这个难题交给机器学习。常见的算法：(1) 人工神经网络[3]，是由大量处理单元经广泛互连而组成的人工网络，用来模拟脑神经系统的结构和功能。神经网络模型相当丰富，典型的有感知器模型，多层向前传播网络，BP 模型，Hopfield 模型，ART 模型等。在模型中，每一个神经元接受多个输入，经过函数与阈值，然后产生输出。(2) 支持向量机[3]，特点是能够同时最小化经验误差和最大化几何边缘。把输入的向量通过核函数映射到一个高纬空间，在这个空间建立一个最大间隔超平面。适合解决非线性的回归，分类等问题。但是缺点也很明显，核函数的确定比较困难。

3 随机森林算法介绍

谈到随机森林，就要谈到组合算法，如图 3 所示。现阶段常使用的组合算法有 2 种。Bagging 方式是让挖掘算法训练的轮数增加，每次的训练样本集合是从初始的训练样本中随机取 x 个样本和 c 个属性出来训练，某个训练样本可能会出现多次。训练后可以得到 r 个弱分类器，如果是分类预测问题，那么采用投票方式综合得出最终预测，如果是回归预测问题，采用简单的平均方式或者是加权平均方式。初始化权重时，常用的方式是，对不同的样本赋予一样的权重，为了公平，权重统一为 $1/n$，接着用算法对该训练样本集训练，使用部分测试数据测试，对于训练失败的样本(分类出错或者是回归值与真实值相差太大) 赋予较大的权重，然后对算法下面的训练集中不好分类的样本再次进行训练，就是把分类出错的样本再次拿来训练，从而得到一组弱分类器，其中每个分类器也是有权重的。最终的预测结果对于分类预测使用投票方式，对回归预测使用加权方式。这两种思想的不同在于，取样的方式不同。但是由于 Boost 要迭代多次，而且 Boost 的各轮训练集需要依赖前一次训练的结果，是一种串行的算法，不能并行化，弱分类器只能按顺序产生。如果对于时间耗费比较多的算法，Boost 就不适合使用。从以上分析来看，文中采用可以并行化的 Bagging 方式。

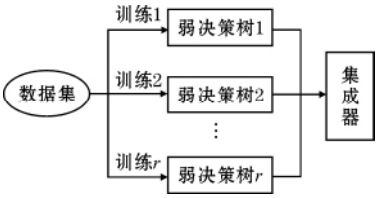

图3　集成学习法结构图

随机森林内部,主要是逻辑回归树,也有 C4.5,对比于任意单个决策树,多树模型效果更好,主要体现在,模型稳定,不容易出现过拟合,准确性也有保障。随机森林算法基于决策树,决策树在研究领域中是一种比较重要的算法模型。决策树中的核心是纯度计算方式,因为要选择最好的分割点,是根据最后的类的纯度来计算,因为是回归预测,所以加入 α 这个阈值,如果在预测值和真实值的误差在 α 范围内,就认为是分类正确,在 α 范围外就认为是分类错误。现在常用的有 4 种纯度计算方式: 基于概率的 Gini 系数,熵,错误率和基于波动的方差。定义为

$$Gini = 1 - \sum_{i=1}^{n} p(i)^2 \tag{2}$$

$$Entropy = -\sum_{i=1}^{n} p(i)\log_2^{p(i)} \tag{3}$$

$$Error = 1 - \max\{p(i) \mid i \in [1,n]\} \tag{4}$$

这些都表示纯度,实际常用的是 *Gini* 系数,文中也采用 *Gini* 做纯度运算。

随机森林的基本原理[6]。弱分类器是逻辑回归树,横向是采用放回抽样的方法得到训练样本集,纵向是采用无放回随机抽样的方式得到训练样本的特征。根据划分,计算纯度,寻得最优切分点,这就是随机森林的基本原理。

随机森林单个弱分类器的训练过程如下: (1) 取样方式使用 Bag,选出每次的训练样本集,假设全训练样本集个数为 N,有回放的抽取样本集 n 来做这个弱分类器的训练。(2) 假设一共有 M 个特征,那么选择 m 个特征作为这个弱分类器的样本特征拿来训练,$m <= M$。(3) 让这棵树完全生长。(4) 输出结果,分类预测用投票,回归预测用均值。

降低误差的方式,在森林中,随机的两个弱分类器的相关度要小,降低此相关度可以减小森林的总体误差率。降低单个分类器的误差,增加准确性也可以提升随机森林的效果。为了减少弱分类器的相关度,使用聚类算法,这里使用 k-means,得到每个训练的训练样本集,因为聚类算法的簇内相似,簇间最大化的间隔,会减少弱分类器之间的相似度。

随机森林分为回归应用与分类应用,通话量预测属于回归应用。随机森林中包含了变量的交互作用,即一个自变量 $X1$ 的变化会导致另一个自变量的 $X2$ 对预测值 Y 的作用发生变化[7]。但是交互作用在其他模型中,如逻辑回归,经常被忽略。在节假日的通话量中,通话量比较离群,但是随机森林由于是多颗树进行组合得出预测模型,对于这种干扰,表现稳定,不容易产生过拟合。

3.1　预测通话量的常用方法

3.1.1　ARMA 预测模型

使用统计学中线性差分方差来描述。常用的是算法模型为自回归滑动平均模型[8-10],常应用于时间序列。模型建立好后,就能用历史数据和现在数据预测出未来的数据。ARMA 模型中有 2 个重要参数 p 自然回归项和 q 滑动平均项,公式为

$$Y_i = u_i + \sum_{i=1}^{p} \varphi_i Y_{i-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{i-1} \tag{5}$$

其中 φ 和 θ 都是模型的参数,u 是一个常数,ε 是误差项。在对原始数据处理后,通过数据的残差序列,得到 ARMA 模型的各个参数值。在处理原始数据时,用相关性来表示两个变量间的相似程度。线性相关,表示两个变量使用线性方程的相关度,非线性相关表示两个变量使用非线性方程的相关度。当前通话量和前一段时间、后一段时间的通话量存在依赖关系,可以使用相关函数来表达。但是效果不好,精度不高,模型不稳定。

3.1.2　SVM(支持向量机) 预测

ARMA 是使用统计分析的方法,是基于大数定律的分析方式,但实际生活中有各种情况,样本数有限,这样 ARMA 难以取得较好的结果。SVM 的基本思想为[11]: 通过核函数,把样本空间从一个低纬的空间经过核函数映射到高维,把非线性的问题,转成高维的线性问题来处理。它不但可以应用于分类预测,还可以用于回归预测。当希望输出的值为类别时,是分类问题,可以分为二分类和多分类,且 SVM 的多分类问题是化分为多个二分类处理的。当希望输出的是连续值时,就是回归问题,它就可以处理通话量的预测。但是找到适合业务场景的核函数比较困难,且模型效果也不是很高。

3.2　随机森林预测通话量过程

3.2.1　构建特征

使用历史的特征来预测将来的通话量,设置要预测的通话量为 y。构造特征的方式如下[12]: (1) 预测当天往前倒推 T 天,计算每天的通话量。(2) 判断预测当天具有的日期信息,如上旬,周末和节假日信息。

(3) 计算预测当前的同比环比。(4) 使用线性回归,得到增益值,也是一个趋势值。(5) 通话中心,当前用户量和用户的一些特征。

因为现在还没有比较好的方法计算每个日期具有的节假日信息,比如春节等,所以日期纬度表是需要人工维护。使用线性回归计算趋势值,主要是根据往前推 T 的每天量,做一元线性回归得到趋势值,公式为[13]

$$b = \frac{\sum_{i=1}^{T} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{T} (x_i - \bar{x})^2} \quad (6)$$

3.2.2 建模过程

设置弱分类器个数为 K,使用 $Gini$ 系数作为类别纯度计算方式。

以前随机森林是随机取每个弱分类的样本,现在改成使用 K-means 聚类算法[14],把相似的样本放在同一个弱分类器中做训练。最后做组合的时候,使用均值组合方式。在模型训练完成后,使用另外一组构建好特征样本,经过模型训练,最后使用精度、召回率、F[15]、方差和误差波动评估模型。

数据来源为平安科技电话中心部分测试数据。图4是日期和通话量的简单二维关系图。

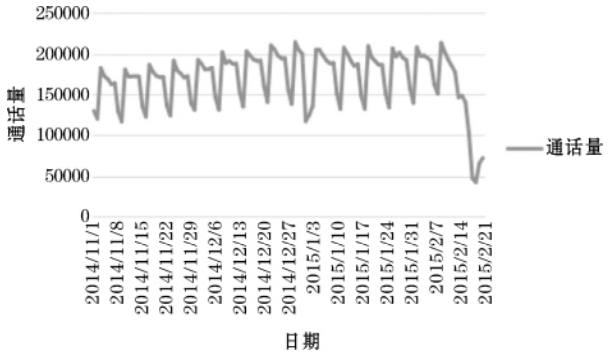

图4 部分通话量与日期关系图

图4简单显示日期和通话量的关系,这只是部分的数据。可以看出有一部分数据是有序稳定的,有一部分数据是跳跃比较大的,且周末数据和节假日数据明显异常。所以用常规的线性回归方式是不能很好地表示规律。

建模过程分为训练和测试2个部分。在训练阶段,主要是根据计算好特征的样本,使用 o-z 归一化,k-means 划分好 k 个弱分类样本后,再进行随机森林训练。在训练完成后,进入预测阶段,测试数据经过刚才训练好的预测模型得到预测值,然后使用真实值和预测值的残差的方差、精度、召回率、F 值来评估模型,最后根据真实值和预测值的误差除以真实值来评估。

3.2.3 结果分析

在训练过程中,分别使用不同簇类个数,来运行多次,其中训练数据是前5年的数据,测试数据是下一个月的数据。多次运行后,选择出测试效果最好的模型。图5是真实值和预测值图。

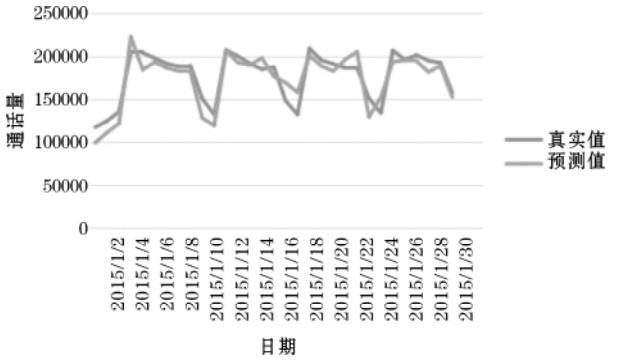

图5 真实值和预测值图

从图5可以看出,在周末的时候,数据残差会稍微大点。测试完后,使用之后几个月的数据验证,方差变化不大,在2000范围内。但是都存在同一个问题,在预测周末与节假日数据时,误差值会比平时大一点,这应该是表示周末和节假日的特征不够完善和周末、节假日的突发性情况有关。

图6是预测数据的误差偏量图。

图6 误差偏量图

误差偏量的平均值为7.1 %,从平均偏差量可以看出,整体预测比较准确。把波动误差在5000内的预测结果归纳为预测正确,5000以外的归纳为预测错误,计算得到精度,召回率,F 值评价指标。各个算法的效果对比如表1所示。

表1 算法效果对比表

算法	评价精度/%	召回率/%	F 值
线性回归	70.23	60.42	0.6496
SVM	72.82	66.65	0.6960
CART	74.94	67.72	0.7115
随机森林	82.62	70.34	0.7599
改进随机森林	87.23	74.28	0.8024

效果显示改进的随机森林算法比以前的算法有明显的提升。尤其是在精度指标上,比没有改进的算法提高了5 %,且随机森林和改进随机森林的 F 值都比以前的单一算法预测效果好。

存在一个问题,可能会出现在某一天,预测值和真实值的误差很高,但是比单独使用线性回归,SVM 的误差低点,可能是因为数据量不够全,或者出现一种新的规律,在模型训练的时候,没有训练到。这种情况出现后,可以重新进行模型训练来解决。还有可能是没有找到可以描述这种情况的特征。

4　结束语

主要对 Spark 和随机森林进行研究,引入 Kmeans 算法作为随机森林的横向划分,引入 Bagging 思想,并使用在预测通话量业务场景中,提升预测精度约5 %。现实情况越来越复杂,很多时间可能会用到在线模型训练,所以需要引入如 Spark 这种分布式计算模型来解决问题。随机森林的功能强大、简单,相信可以应用在更多的领域。

参考文献:

[1] 程浩,黄杰. Spark 大数据处理技术 [M]. 北京: 电子工业出版社,2015.

[2] 耿嘉安. 深入理解 Spark: 核心思想与源码分析 [M]. 北京: 机械工业出版社,2015.

[3] 王晓华. Spark MLlib 机器学习实践 [M]. 北京: 清华大学出版社,2015.

[4] 陈蓉. 话务量分析和多种预测模型的比较研究 [M]. 北京: 北京邮电大学,2008.

[5] 蒋盛益,李霞,郑琪. 数据挖掘原理与实践 [M]. 北京: 电子工业出版社,2011.

[6] Yunming Ye,Qingyao Wu,Joshua Zhexue Huang. Stratified Sampling For Feature SubSpace Selection In Random Forests For High Dimensional data [J]. Pattern Recognition,2013,46(3): 769 – 787.

[7] 李欣海. 随机森林模型在分类与回归分析中的应用 [J]. 应用昆虫学报,2013,50(4): 1190 – 1197.

[8] George Box, Gwilym M,Jenkins et al. Time Series Analysis: Forecasting And Control [J]. Journal of the American Statistical Association, 2016, 68(342): 343 – 344.

[9] Mills,Terence C. Time Series Techniques For Economists [M]. Cambridge University Press,1990.

[10] Percival,Donald B,Andrew T Walden. Spectral Analysis For Physical Applications [M]. Cambridge University Press,1993.

[11] Limeng Cui,Yong Shi. A Method Based On One-class SVM For News Recommendation [J]. Procedia Computer Science,2014,31: 281 – 290.

[12] 曹国,沈利香,基于案例推理的银行零售客户价值细分模型构建 [J]. 财会月刊: 理论版,2011,33: 18 – 22.

[13] Zhizheng Liang, Youfu LimShiXiong Xia. Adaptive Weighted Learning For Linear Regression Problems Via KullBack-Leibler Divergence [J]. Pattern Recognition,2013,46(4): 1209 – 1219.

[14] Shigei, Miyajima, Maeda. Bagging and AdaBoost algorithms for vector quantization [J]. Quantization, Neurocomputing, 2009, 73(1 – 3): 106 – 114.

[15] 陈沛玲. 决策树分类算法优化研究 [D]. 长沙: 中南大学,2007.

Call Prediction Research based on Random Forest and Spark Calculation Model

WANG Qi,　ZHANG Hong-wei

(College of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: In order to improve the prediction effect of call and improve the recall rate of model prediction, the precision and the F, it uses Random Forests model based on multiple decision trees, using a combination of Bagging way and multiple weak classifiers. Random Forests in Spark is easily parallelized. Spark than MapReduce is more suitable for iterative arithmetic and interactive mining. It raises prediction in the center of the peace of telephone calls. The Random Forests algorithm based on combination of Bagging will definitely enhance the effect of numerical model forecast.

Key words: computer application; intelligent engineering; Spark; Hadoop; call prediction; random forest