公交网络中的乘客需求预测系统和方法

周春姐 张志旺 唐文静
(鲁东大学信息与电气工程学院 山东 烟台 264000)

摘 要 公共交通工具,尤其是公交车服务,可以减少私家车的使用和燃油消耗,缓解交通拥堵和环境污染状况。当乘坐公交车时,乘客不仅关心等车时间,更在乎公交车的拥挤程度,过度拥挤的公交车会导致乘客放弃乘坐。可见,准确、实时、可靠的乘客需求预测可以帮助公交公司决定合理的公交发车时间间隔,并且可以减少乘客的等车时间,这是人们急切需要的。基于实际公交系统的大量数据,提出一个面向移动用户的乘客需求预测系统。该系统包括服务器端的信息数据流处理和挖掘程序,以及客户端的移动应用程序。然而,公交网络中的乘客需求预测存在三大挑战:不均匀性、突发性和周期性。为了解决这些问题,提出了 3 种预测模型和 1 种基于滑动窗口的框架来预测乘客的数目。开发了一个原型系统,该系统可运行在多个版本的 Android 移动手机上,22 个月的连续实验证明,该系统能够对公交网络中的 864110 项乘客需求进行精确预测,其准确度超过 78%。

关键词 公交车服务,交通拥堵,乘客需求预测,预测模型
中图法分类号 TP391 **文献标识码** A

System and Methods of Passenger Demand Prediction on Bus Network

ZHOU Chun-jie ZHANG Zhi-wang TANG Wen-jing
(Department of Information and Electrical Engineering, Ludong University, Yantai, Shandong 264000, China)

Abstract Public transport, especially bus transport, can reduce the private car usage and fuel consumption, and alleviate the condition of traffic congestion and environmental pullution. However, when traveling with buses, the travelers not only care about the waiting time, but also care about the crowdedness in the bus. Excessively overcrowded buses may drive away many travelers and make them reluctant to take buses. So accurate, real-time and reliable passenger demand prediction becomes necessary, which can help determine the bus headway and reduce the waiting time of passengers. However, there are three major challenges for predicting the passenger demand on bus services: inhomogeneous, seasonal bursty periods and periodicities. To overcome the challenges, this paper proposed three predictive models and further took a data stream ensemble framework to predict the number of passengers. This paper developed an experiment over a 22-weekperiod. The evaluation results suggest that the proposed method achieves outstanding prediction accuracy among 86411 passenger demands on bus services, more than 78% of them are accurately forecasted.

Keywords Bus transport, Traffic congestion, Passenger demand prediction, Predictive models

1 引言

随着经济的高速发展,交通运输业也得到了飞速发展,但是交通状况却不断恶化,产生了一系列交通问题:交通拥挤和道路阻塞的现象日趋严重,交通事故频繁发生,随之而来的能源消耗和环境污染也越来越引起社会的普遍关注。公共交通工具,尤其是公交车服务可以有效缓解这些问题。公交服务能有效地利用现有交通设施,减少交通负荷和环境污染,保证交通安全,提高运输效率,改善道路使用者的方便性和舒适性。另外,公交车分布范围广,价格便宜,日益受到人们的青睐。2011 年新加坡的 500 万居民中平均每天乘坐公交车的就有 330 多万。然而,目前公交车的服务质量还有待提高,乘客在乘坐公交车时希望尽可能缩短等车时间和乘坐不拥挤的公交车。事实上,过度拥挤的公交车可能会吓走很多乘客,从而使他们放弃乘坐公交车。因此,合理均衡的公交服务才能使公交公司和乘客双方的利益最大化。如果失去这种均衡,就会出现以下情形:1)过多的空车和很少的乘客需求;2)乘客的等待时间长和过度拥挤的公交车。可见,准确、实时的乘客需求预测可以帮助公交公司决定合理的公交发车时间间隔,并且可以减少乘客的等车时间,这正是人们急切需要的。

然而,由于很多不确定因素的存在,本研究面临如下 3 个挑战:1)非均匀性。乘客对公交服务的需求在不同站点、不同工作日,以及同一天的不同时间段都存在差异。2)突发性。每个公交站点的乘客需求量不同,且很多公交站点的乘客需

本文受国家自然科学基金项目(61202111,61472141,61273152,61303017),山东省高等学校科技计划项目(J12LN05),山东省自然科学基金联合专项项目(ZR2013FL009),烟台市科技发展计划项目(2013ZH092,2014JH042),鲁东大学博士基金项目(LY2012023)资助。
周春姐(1981-),女,博士,副教授,主要研究方向为物联网、数据挖掘,E-mail:lucyzcj@gmail.com;**张志旺**(1979-),男,博士,副教授,主要研究方向为数据挖掘与知识发现、机器学习、人工智能和自然语言处理;**唐文静**(1980-),女,博士,副教授,主要研究方向为数据融合、模式识别、图像分析与处理。

求存在突发性,会受到很多意外事件的影响,如交通拥堵、天气变化等。3)周期性。乘客对公交服务的需求在不同周的同一工作日,以及同一天的早晨和傍晚都存在很高的相关度。为了很好地解决这些挑战性问题,本文提出了一种基于公交网络的乘客需求预测系统和方法。基于历史的 GPS 数据和公交服务数据(如乘客的上/下车站点),针对每个公交站点的乘客需求建立一个 P 分钟的时间序列直方图。本研究采用有名的时间序列预测技术,如时变泊松模型、加权时变泊松模型和综合自回归移动平均模型等来预测公交网络中的乘客需求量。

准确、实时的乘客需求预测可以使公交公司和乘客双方受益。为了帮助乘客做出明智的出行决定,本文提出了一个系统,可以实时地向乘客手机推送当前的公交状态和交通信息。该系统分析预测了在公交站点等候的乘客数目、即将经过该公交站点的乘客数量,以及公交站点附近区域的交通状况。通过掌握公交车和交通状况的实时信息,乘客可以做出明智的决策。

该系统是可行的,因为目前大多数公交车都拥有 GPS 设备,并且大部分乘客都有手机。本研究利用了真实的公交网络数据集,该数据集包含了烟台市 416 个公交站点,1326 辆公交车的信息。我们的测试床是一个线下运行的计算流模拟。前 16 个月的数据被用作训练集,后 6 个月的数据被用作流式实验台的输入,即模拟实验流中不断到达的乘客需求。

本文的主要贡献如下:

1)构建了面向移动终端用户的乘客需求预测系统。该系统包括服务器端的信息数据流处理和挖掘程序,以及客户端的移动应用程序。

2)提出了 3 种预测模型和 1 种基于滑动窗口的框架来预测乘客的数目和公交车的拥挤程度,从而便于移动终端用户做出个性化的选择。

3)提出了知识库的设计方法,包括其物理表示模型以及知识和规则的相关分析。

4)提出了一种停靠点的检测方法。在这些停靠点,公交车因为要搭乘乘客或者司机需要休息而逗留了一段时间。首先利用基于密度的聚类算法得到停靠点候选集,然后利用一种监管模型滤除由于交通拥堵或者交通信号灯导致的错误候选集。

5)使用了 1326 辆公交车在 22 个月内产生的真实数据集,并且在多个版本的 Android 移动手机上评估该系统,实验结果表明,该方法能高效地预测公交网络中的乘客需求。

本文第 2 节介绍了相关研究工作;第 3 节给出了基于云的预测系统;第 4 节提出了 4 种预测模型;第 5 节提出了知识库相关规则;第 6 节介绍了移动终端应用;第 7 节利用真实数据集通过实验验证了所提算法的有效性;最后总结全文并展望未来。

2 相关工作

目前,很少有研究关注公交网络中的乘客需求预测。相关工作可以概括为以下 3 个方面:

1)发车间隔决策。Vuchic 等提出一种最大负荷区的方法来决定公交车的发车时间间隔,从而提供足够的交通运输能力。Daganzo 等[1]提出一种自适应控制模式来消除总线聚束,根据实时的车头时距信息来动态地决定公交车的发车时间。Yan 等[2]针对城际巴士线的随机需求研究了路由和时间表的设置模型,然而这种模型没有考虑市内公交线路,也没有分析不同时间段的需求差异。

2)公交车到站时间预测。掌握准确、实时、可靠的公交车到站时间信息对公交公司和乘客都是有益的。然而,由于一些可变因素的影响(例如,交通拥堵、天气条件和路口延迟等),公交车到达时间的精确预测是具有挑战性的。Chang 等[3]基于最近邻非参数回归提出了一种动态模型来预测从起始站点到终结站点的多条路径行驶时间。Yu 等[4]提出了包括支持向量机(SVM)、人工神经网络(ANN)、K 近邻算法(KNN)和线性回归(LR)等多种方法来预测等待时间,结果证明 SVM 模型对多条线路的公交站点的等待时间预测是最好的。

3)交通拥堵。关于交通堵塞目前也有相关的研究工作。在 StreetSmart[5]中,每辆车绘制一个基于附近其他车辆速度的地图,并且将其发送给相邻的车辆。进而,文献[6]中计算了相邻车辆的平均速度。文献[7]提出了一个车载通信系统,通过挖掘和传播道路信息实现交通堵塞的检测和预警。该系统由两部分组成:1)基于泛洪的协议,用于传输交通信息;2)迪杰斯特拉算法,用于动态地计算车辆的最少拥堵路线。文献[8]设计了一种虚拟的交通灯协议,在不需要任何路边基础设施的情况下,对道路交叉口的运输流量进行动态优化。

上述方法都假设道路中的每辆车自愿参与交通管理,并且主动提供相关信息。但实际上,很多司机只愿意享受交通信息带来的便捷性,而不愿意分享任何信息。由于这种自私行为的存在,这些交通管理系统都是不可行的。本文的乘客需求预测可以帮助公交公司决定合理的公交发车时间间隔,减少乘客的等车时间,从而减少甚至避免交通堵塞。

3 基于云的预测系统

本文提出了一个基于云的预测系统,用于为移动终端用户预测公交网络中的乘客需求量。系统中利用了大量配备 GPS 的公交车和具有 GPS 功能的乘客移动手机。系统框架如图 1 所示。

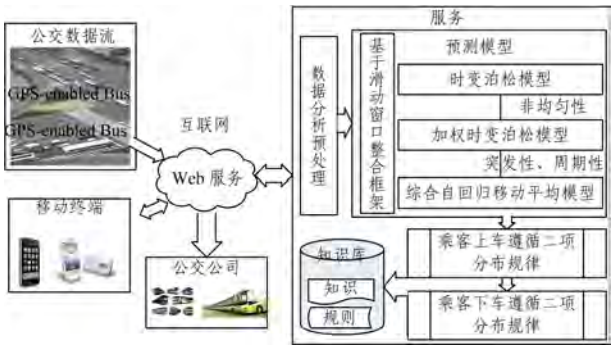


图 1 系统框架

该系统包括 3 个主要部分:公交 GPS 数据流、服务器(基于云)和移动客户端。公交车经由无线通信将数据流发送给服务器。服务器端的应用程序处理该数据流,并不断更新公交站点的乘客需求量。服务器端向公交公司提供乘客对公交服务的需求,同时为移动终端用户提供接口,用于检索公交车信息和交通信息。将运行在乘客手机上的移动应用程序连接

到服务器,并根据乘客所在位置检索相关的信息。本节将描述公交数据流和一些服务器端程序。乘客需求预测模型将在第 4 节介绍,第 6 节介绍移动终端应用。

3.1　公交数据流

本文采用的数据集包含 1326 辆公交车,806257 个上/下车站点的公交数据,全面地覆盖了整个烟台市区的、22 个月的公交信息。根据公交车的换班时间,将运行时间分为 4 个时间段(早晨 5 点到上午 9 点,上午 9 点到中午 1 点,中午 1 点到下午 5 点,下午 5 点到晚上 9 点)。公交数据包括 5 个属性值:1)公交状态值,其中 busy 表示乘客的数量大于公交车的容量,free 表示乘客的数量小于公交车的容量,park 表示公交车正停靠在起始或终结站点上;2)公交站点的 ID;3)数据产生时间;4)公交车牌号;5)GPS 数据对应位置的经纬度。

3.2　数据预处理和数据分析

"数据预处理和数据分析"对从"公交 GPS 数据流"传来的源数据进行预处理,包括消除噪声和滤除与乘客需求预测无关的冗余数据,并且利用统计等方法对公交数据进行预分类。

3.3　停靠点的检测方法

停靠点定义为公交车因为需要搭乘乘客或者司机需要休息(如吃饭、喝水或者去洗手间等)而逗留了一段时间的地点。停靠点的检测很重要的原因在于它不同于行车状态,也会有很多附加应用,如某个城市的热点检测等。首先利用基于密度的聚类算法得到停靠点候选集,如图 2(a)－图 2(d)所示;然后利用一种监管模型滤除掉由于交通拥堵或者交通信号灯导致的错误候选集,如图 2(e)－图 2(f)所示。

停靠状态的一个重要特征是它跨距了一系列非常相近的 GPS 节点。因此,首先利用候选集方法来选取这些局部的节点集合。对于轨迹 $TR=\{P_k\}_{k=1}^{n}$,如果满足条件 $dist(P_i,P_{i+j})<\delta(1\leqslant j\leqslant m)$ 和 $timeDiff(P_{i+m},P_i)>\tau$(其中,$P_i$ 是枢点,m 是满足条件的最大数目,δ 和 τ 是阈值参数),则得到停靠点的一个候选集 $PR\{P_i,\cdots,P_{i+m}\}$。得到这个候选集之后,再将 P_{i+1} 作为下一个枢点(见图 2(c)和图 2(d))来扩充候选集,直到上述两个约束条件不成立。这种逐渐增长的候选算法是广度优先搜索算法和基于密度聚类算法的融合。重复该过程,直到轨迹中的所有节点都被扫描完,得到轨迹 TR 的停靠点候选集 $\{PR_i\}_{i=1}^{K}$。

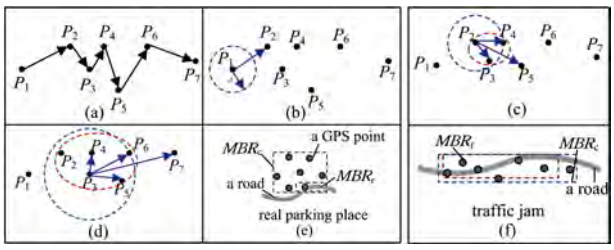


图 2　停靠点检测方法示例

从本质上讲,该候选集检测算法是在空间、时间和速度约束条件下,查找公交车 GPS 节点的密集区域。然而,这些候选的停靠点有时是由于交通拥堵或者交通信号灯引起的,并非是真正的停靠。

为了减少此类错误候选集,本文采用一种监管模型用于从候选集中选出真正的停靠点。该模型具有如下特性:1)最小边界化(Minimum Bounding Ratio,MBR),如图 2(d)－图 2(e)所示。*MBR* 表示路段边界框(*MBRr*)和候选集中 GPS 节点边界框(*MBRc*)之间的比率;2)平均距离,指候选集中各节点与离它们最近的路段之间的距离平均值;3)中心距离,指候选集 *MBRc* 的中心节点到路段的距离;4)时间段,指停靠的时间段 t;5)历史值,指过去的 7 天 50 米范围内停靠点的数目;6)poi 矢量,指 50 米范围内某特定 POI 的数目。

4　预测模型

本文的目标是预测 t 时刻在公交站点 s 需要乘坐 b 路公交车的乘客数目。为此,提出了 3 种预测模型和 1 种基于滑动窗口的框架。

4.1　时变泊松模型

给定时间内某公交站点有 n 辆公交车停靠的概率 $P(n)$ 满足泊松分布,定义为:

$$P(n;\lambda)=\frac{e^{-\lambda}\lambda^{n}}{n!}$$

其中,λ 表示在固定时间段内乘客对公交服务的平均需求比率,λ 的值随时间变化。因此将其看作一个时间函数 $\lambda(t)$,从而将泊松分布变换成非齐次的。$\lambda(t)$的定义为:

$$\lambda(t)=\lambda_1\delta_{d(t)}\eta_{d(t),h(t)}$$

其中,$d(t)$表示工作日{1＝周日,2＝周一,…};$h(t)$是时间 t 所属的时间段(例如,若将每 30 分钟作为一个时间段,则时间 00:31包含于第二个时间段)。另外,需要满足下面两个等式:

$$\sum_{j=1}^{7}\delta_i=7,\ \sum_{D=1}^{D}\delta_{D,i}=D$$

其中,D 是一天中时间段的数目;λ_0 是一周泊松过程的平均比率;δ_i 表示第 i 天的相对变化(如周六的比率低于周二);$\eta_{j,i}$表示第 j 天第 i 时间段的相对变化(如高峰期);$\lambda(t)$是一个离散函数,用于表示公交站点 s 中随时间变化的乘客需求。

4.2　加权时变泊松模型

时变泊松模型只预测了时间相关的平均乘客需求,然而每个公交站点的乘客需求量是不同的。实际上,很多公交站点的乘客需求具有突发性,受到很多意外事件的影响,如交通拥堵、天气变化等。加权时变泊松模型能很好地解决突发性问题,其目的是增加上周乘客需求量与前几周乘客需求量的相关度[9]。相关度的权值 w 采用有名的时间序列方法——指数平滑法来计算,其定义为:

$$w=\alpha*\{1,(1-\alpha),(1-\alpha)^2,\cdots,(1-\alpha)^{\lambda-1}\}$$

其中,λ 是以往时间段中乘客需求的平均值;α 是平滑因子,其值是由用户定义的,取值范围为 $0<\alpha<1$。

4.3　综合自回归移动平均模型

上述两种模型都假设乘客对公交服务的需求存在周期规律性,而实际上乘客的需求在不同站点、不同工作日以及同一天的不同时间段都存在差异。综合自回归移动平均模型可以很好地模拟和预测单变量时间序列数据,如交通流数据和短期预测问题等。其优势在于能够准确地表示不同类型的时间序列,如自回归时间序列、移动平均时间序列以及二者的结合。在综合自回归移动平均模型中,变量的预测值可以看作历史观测和随机误差的线性函数。本文将某特定公交站点 s 上随时间变化的乘客需求量看作时间序列,预测过程可以表示为:

$$R_{s,t}=\theta_0+\phi_1X_{s,t-1}+\phi_2X_{s,t-2}+\cdots+\phi_pX_{s,t-p}+\varepsilon_{s,t}-\theta_1X_{s,t-1}-\theta_2X_{s,t-2}-\cdots-\theta_qX_{s,t-q}$$

其中，$R_{s,t}$和$\varepsilon_{s,t}$分别是在时刻t乘客需求量的实际值和随机误差；$\phi_l(l=1,2,p)$和$\theta_m(m=0,1,2,\cdots,q)$是模型的参数和权值，其中$p$和$q$表示模型的阶的正整数。模型的阶和权值都可以利用自相关函数和偏自相关函数从历史时间序列中得到。这些值可以用来检测是否存在周期性及其周期性的频率。

4.4 基于滑动窗口的整合框架

上述3种预测模型分别针对长期、中期以及短期的历史数据进行预测。基于滑动窗口的整合框架旨在将它们结合起来以实现更好的预测。$M=\{M_1,M_2,\cdots,M_z\}$表示对一个给定时间序列进行建模的z个模型的集合；$M_t=\{M_{1t},M_{2t},\cdots,M_{zt}\}$表示这些模型在时刻$t$对下一时间段的预测值的集合。整合预测值$E_t$的计算式如下：

$$E_t=\sum_{i=1}^{z}\frac{M_{it}}{\beta},\beta=\sum_{i=1}^{z}\rho_{iH}$$

其中，ρ_{iH}是模型M_i在时间窗口$[t-H,t]$内的某个时间段预测做出的值，H是由用户定义的滑动窗口的大小。因为在后续的时间段中公交数据信息持续到来，因此时间窗口也要不断滑动，从而保证这些模型在下一个H时间段内正常运行。为了更好地评价预测的准确性，采用著名的时间序列预测误差度量机制－对称平均百分比误差(sMAPE)[10]。

5 知识库

知识库中存储经过需求预测分析后的各类规则和知识。在此之前，首先对需求预测分析步骤中检测出的结果进行评估。经过用户或机器评估后，可能会发现其中存在冗余或无关的结果，此时应该将其剔除。知识库中只保留那些经过评估和验证后的、能真实反映乘客需求的、有用的知识和规则。这些知识被存储在云端，用于更好地服务于乘客和公交公司。

5.1 物理表示模型

本文提出了一种基于树的模型用于表示公交车和乘客的关系。其中叶子节点存储每路公交车；当乘客到达时，他们的信息被存储在不同的叶子节点中。例如，若某乘客乘坐的是公交车Bus1，则该乘客的上下车时间就存储在相应的叶子节点中。中间节点存储的是公交车发车时间间隔关系，每个中间节点都对应一系列谓词(如之前、交叉、平行等)。图3(a)展示了一个树模型结构，该结构是一个左深树，Bus1和Parallel首先结合，其输出结果再与Bus4匹配；同理，Parallel和Bus4首先结合，然后再与Bus1匹配的右深树也是可行的。

算法1 栈的存储模式

```
Input: N: the number of passengers in the stack; SP: the get-on time of
       the passenger; EP: the get-off time of the passenger; k: one of
       the passengers that already in the bus's stack
Output: The correct TList of the stack; string[ ] A
1. for i=0; i < N; i++  do
2.   for j=1; j<=N; j++  do
3.     if A[i]. EP has arrived and there is no k satisfying the prior
       pointer of A[k] is A[j]. SP  then
4.       the prior pointer of A[j]. SP is set to A[i]. EP;
5.       the prior pointer of A[j]. EP is set to A[j]. SP;
6.     else if A[i]. EP has arrived and there is a k satisfying the prior
       pointer of A[k] is A[j]. SP  then
7.       the prior pointer of A[j]. SP is set to A[i]. EP;
8.       the prior pointer of A[j]. EP is set to A[k];
9.     else if A[i]. EP has not arrived and there is no k satisfying the
       prior pointer of A[k] is A[j]. SP  then
10.      the prior pointer of A[j]. SP is set to A[i]. SP;
11.      the prior pointer of A[j]. EP is set to A[j]. SP;
12.    else
13.      the prior pointer of A[j]. SP is set to A[i]. SP;
14.      the prior pointer of A[j]. EP is set to A[k];
15.    end if
16. end for
17. end for
```

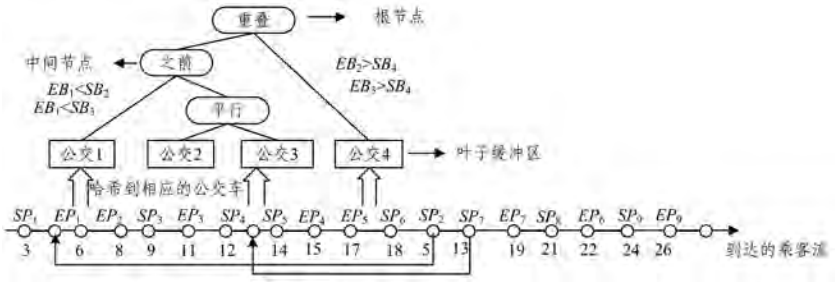
算法2 栈的插入操作

```
Input: N: the current number of passengers in the bus stack; SP: the
       get-on time of the passenger; EP: the get-off time of the passen-
       ger; k: the new arriving passenger; MS: the maximal size of the
       stack
Output: 'he correct TList of the stack
1. while k<MS do
2.   if all records in the stack are in order then
3.       the prior pointer of A[k]. SP is set to A[N]. EP;
4.       the prior pointer of A[k]. EP is set to A[k]. SP;
5.     else if there are out-of-order records in the stack, and the get-
       off time of all passengers have arrived then
6.       for i=N-1; i>=0; i--  do
7.         if A[k]. SP>A[i]. SP  then;
8.           the prior pointer of A[k]. SP is set to A[i]. SP;
9.         else
10.          the prior pointer of A[k]. SP is set to A[i]. EP;
11.          the prior pointer of A[k]. EP is set to A[k]. SP;
12.        end if
13.      end for
14.      else
15.        wait until there is no absent get-off time;
16.    end if
17. end while
```

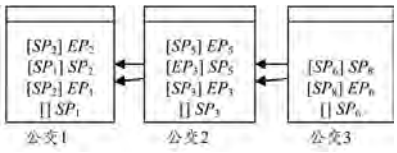
树中的每个节点都有一个栈，分别存储到达的乘客信息(对于叶子节点)和换车的乘客信息(对于中间节点)。每个栈都包含一些记录值，每条记录都有一个乘客指针，分别指向该乘客的上车和下车时间。对于栈中每个乘客的上车时间，都有一个额外的值PreEve用于记录前一个状态栈中的、按照时间序列排列的最邻近的一个乘客(见算法1)。对于每个乘客的下车时间，PreEve首先指向其对应的上车时间。当其上车时间成为其他乘客的PreEve时，其PreEve就改为指向该乘客。例如，图3(b)中，EP_1的PreEve首先设定为SP_1。在栈Bus1中距离SP_2最近的乘客是SP_1。那么SP_2的PreEve就指向SP_1，并且SP_1的PreEve改为指向SP_2。中间节点的开始和结束时间是与该节点相关的所有公交车的最早开始时间和最晚结束时间。

图3(b)表示树模型中每个节点的栈存储。在每个栈中，各个实例按照其到达的时间顺序自上而下存储(见算法2)。对于有序记录，每个到达的乘客信息只是简单地存放于相应栈的底部，其PreEve指向前面栈的最后一个乘客。然而，实际操作中可能会出现乱序记录，具体介绍请参考文献[11]。此时，这种简单的添加措施将不适用于乱序事件的插入。一

条乱序记录被存放于相应栈中,并且按照到达时间排序。如果某趟公交车上的所有乘客的下车时间都已经到达,并且乘客 k 的上车时间 $SP_k > SP_i(EP_i)$,那么 SP_k 的 PreEve 将指向 $SP_i(EP_i)$;否则,将等待缺失的下车时间。如果其上车时间已经到达,那么乘客 k 的下车时间 EP_k 的 PreEve 将指向其对应的开始时间 SP_k。例如,图 3(a)中,假设存在乱序记录 SP_2 和 SP_7。在 EP_7 到达之前,SP_7 的 PreEve 不能被设置为 SP_4(此时,乘客 4 和乘客 7 乘坐同一辆公交车),因为在时刻 19 之前,只收到乘客 7 的上车时间,但没有下车时间。同理,SP_2 的 PreEve 设置为 SP_1,因为其对应的下车时间已经到达。



(a)基于树的模型



(b)叶子缓冲区存储

图 3　物理表示模型

5.2　知识与规则的相关分析

通过 4.1 小节中时变泊松模型的分析可以看出,乘客每天对公交服务的需求有一定的规律性。图 4 给出了乘客某个月对公交服务的需求量分析,反映了乘客在该段时间内的行为模式,其数据呈非均匀性。

图 4　一个月的数据分析结果

通过 4.2 小节中加权时变泊松模型的分析可以看出,许多公交站点的乘客需求受突发事件的影响而呈现出季节性,如造成交通拥堵的节假日、天气变化等,如图 5 所示。本文的目标就是增加本周与前几周需求模式的相关性。例如,上周一的状况与两周或者三周前周一的状况的相似度很高,如图 5 中两个日常的周一。同样是周一,日常和下雨天存在明显差异,其相关性明显降低,如图 5 中一个日常的周一和一个雨天的周一。

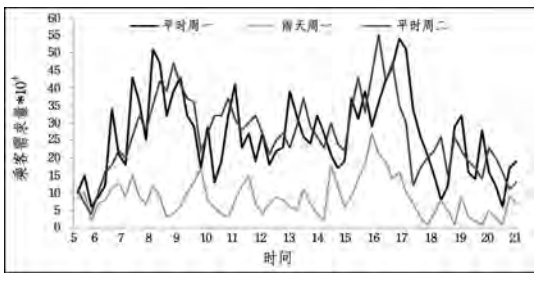


图 5　每天的乘客需求(日常和雨天的对比)

通过 4.3 小节中综合自回归移动平均模型的分析可以看出,正常工作日中,在某个公交站点,乘客对某趟公交车的需求量相似度很高(如周二和周三);并且,乘客对公交服务的需求在同一天的早晨(如上午 7 点到 9 点的上班时间)和傍晚(如下午 5 点到 7 点的下班时间)也很相似。图 6 给出了某繁忙公交站点上乘客对公交服务需求的时间序列。我们研究了有关乘客需求的 16 个月的数据的自相关曲线,其中每 12 小时就有一个相关高峰。

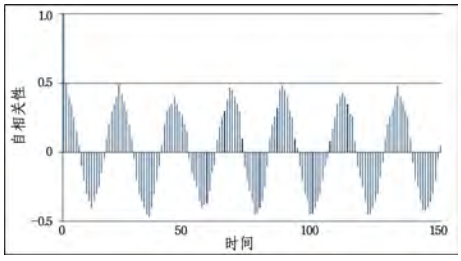


图 6　乘客需求的自相关性

6　移动终端应用

为了方便移动终端用户获取公交车的状态,我们搭建了一个基于位置的、运行在智能手机上的移动应用程序。图 7 给出了移动应用程序运行在 Android 手机上的一些截图。我们提供了一个友好的用户界面,允许用户方便地操作不同的功能。此外,大多数功能都能通过谷歌地图直观地显示出来,从而有助于用户更好地查看结果。

通过分析该移动应用程序能够给出适合某乘客出行的公交车列表,同时乘客也可以输入一个地址来检索该地址附近的公交站点,如图 7(a)所示。应用程序可以在地图上显示所检索的公交站点,如图 7(b)所示。乘客的当前位置在地图上用气球标注。当乘客点击某个公交站点时,将要在该站点停靠的附近所有公交车被显示出来,如图 7(c)所示。当乘客选

中某趟具体的公交车时，该公交车的状态信息就会由服务器传给乘客，如图 7(d)所示。公交车的到达时间、公共车上的乘客数量、公交站点等待的乘客数目、未来几分钟内将要经过该站点的乘客数目、公交车的状态以及当前的行车速度都会呈现给乘客。乘客基于这些信息来决定是否乘坐公交车。

(a) (b) (c) (d)

图 7 移动终端应用截图

运行在用户 GPS 手机上的移动客户端接收用户的查询，与云端进行通信，并将结果呈现给用户。移动客户端通过用户的出行习惯逐渐学习用户的出行模式，且持云端定制适合该用户的可行的、最快的出行路径。

7 实验结果与分析

本节首先描述了实验的建立方法，使其能够利用现有数据更好地测试所提模型，然后讨论了相关实验结果。

7.1 实验的建立

所提模型每 P 分钟对所有公交站点的乘客需求进行一次在线预测。这种测试是通过一个线下的连续仿真来实现的，其中用到了 R 统计软件。

将数据集分成两部分，其中 16 个月的数据作为训练集，6 个月的数据作为测试集。训练集和测试集均由每个公交站点的时间序列组成，并且时刻 t 的每个测试值都被加入到训练集中用于预测时刻 $t+1$ 的值。例如，30 分钟内某公交站点上每辆公交车的乘客需求量可以用于预测下一个 30 分钟的乘客需求量，以此类推。

ARIMA 模型通过学习和检测运行在每个公交站点的历史时间序列上的自相关和偏相关函数得到，并且每 24 小时更新一次。为此，本文利用了 R 包中的自动时间序列函数-auto-arima，其中的参数是默认的；同时利用了 R 包中的 arima 函数，其中每个时间段的权值和参数都有其确定的值。

无论带权值的还是无权值的时变泊松模型均每 24 小时更新一次。在整合模型中滑动窗口被设定为 4 个小时。我们采用公制来衡量每种模型的准确性，这种方法同样适应于整合框架中的每种模型——sMAPE。

因为没有非常相关的研究工作来预测公交系统中的乘客需求，所以为了证明本文模型的有效性，我们与出租车预测领域中的一些相似方法进行了比较，同时还比较了一种平均预测方法 APM。APM 是通过随机预测 300 次乘客的需求量，然后取其平均值来实现的。所有模型的报告值都是每个时间序列中计算出的值的平均。

7.2 实验结果

我们对 86411 项公交服务进行了测试。表 1 列出了一天中各个公交站点的乘客需求量。从表中可以看到，工作日的总公交服务量多于周末的。这是因为在大多数情况下工作日有更多的乘客出行。此外，在工作日中，上午 5 点到上午 9 点和下午 5 点到下午 9 点这两个时间段内的平均乘客需求超过了时间段上午 9 点到下午 1 点和下午 1 点到下午 5 点。而在周末，时间段上午 5 点到上午 9 点内的平均乘客需求显著减少，但下午 5 点到下午 9 点这个时间段内的仍然很多。这是因为人们在周末不上班，但他们晚上仍会有一些休闲娱乐活动。

表 1 公交服务(每天)

	总公交服务量	平均乘客需求量			
		上午 5 点到上午 9 点	上午 9 点到下午 1 点	下午 1 点到下午 5 点	下午 5 点到下午 9 点
工作日	1326	73200	41600	39800	67500
周末	982	28600	36400	21900	43700
总量	2308	101800	78000	61700	111200

表 2 列出了某趟公交车的所有服务信息。表 2 中的“需求”表示已经上车的乘客数量；第二列与公交车之间的发车时间间隔有关，并且这些公交车的状态均为“忙”。由表 2 可以看出，大约 24% 的公交车处于“空闲”状态，而 76% 的处于“忙”状态；乘客从上车到下车的平均时间间隔是 21 分 12 秒。

表 2 公交服务(每辆车)

	需求	处于“忙”状态的时间/h
最大	893600	9
最小	9200	2
总平均	24700	3
平均 $t<30$ m	21300	

图 8 和图 9 分别给出了工作日和周末停靠点的分布概览。从两个图可以发现，热点区域的分布具有一定规律性，并且工作日和周末的分布有差异性。例如，有些区域是持续热点，只是其范围随时间有所变化；而有些区域的热点度与时间息息相关。进而，我们还可以发现在工作日人们的活动区域集中在工作场所和居民区，而周末则集中在娱乐场所和学校。

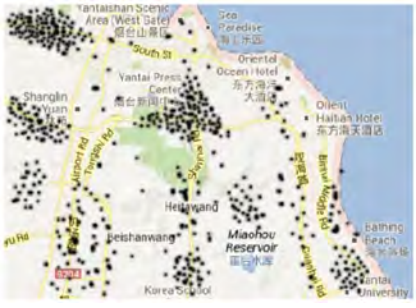


图 8 工作日的停靠点分布

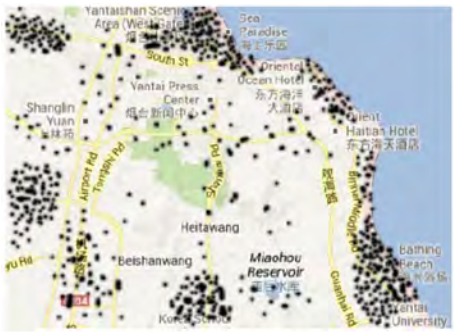


图 9　周末的停靠点分布

图 10 对文献[12]中的 4 种算法进行了测试，其中层次聚类使用的是平均连锁模型。当截止距离设定为 500 米时，图 10(a)中产生了 70 个聚类。在 DBSCAN 算法中，当密度阈值设定为在半径为 200 m 的范围内设 10 个点时，图 10(b)中的所有记录被分成 7 个聚类。对于 K-means 算法，图 10(c)和图 10(d)分别显示了 70 个聚类和 7 个聚类的情况。文献[12]分析发现不同的数据分布直接影响着这种聚类算法的性能，很难给出判断标准，因此急需一种全新的算法来整合这些聚类算法的优点。本文所提出的停靠点检测方法能很好地解决这一问题；并且，该方法还能够滤除由于交通堵塞或者交通信号灯导致的错误结果，而这些情况在文献[12]中没有被考虑到。文献[12]的未来工作中提出某些著名的事件也有可能影响出租车需求量的分布。长远来看，本文所提知识库能够在一定程度上解决这一问题。

(a) AHC using Average Linkage at 500 m

(b) DBSCAN(Eps=200 m and MinPts=10)

(c) K-means(k=70)

(d) K-means(k=7)

图 10　不同聚类算法产生的聚类结果

图 11 显示随着乘客数量的增多，其平均等待时间的变化趋势。由图可知，变化趋势可划分为 3 个阶段。当乘客数量少于 20 时，平均等待时间随着乘客数量的增多而迅速增加；在乘客数量增加到 60 之前，平均等待时间的增长速度趋于缓慢；随后，其增长速度再次加快。原因在于，每辆公交车大概有 20 个乘客座位，而有 40 位乘客可以站在公交车上。当乘客数量少于 20 时，乘客越多就占用越多的上下车时间。因此，这种情况下乘客数量对平均等待时间的影响很大。当乘客数量为 20～60 时，乘客可能要花费更多的上下车时间，但是它对平均等待时间的影响较小。当乘客数量大于 60 时，乘客需要等待下一辆公交车，因此平均等待时间的增长速度再次加快。

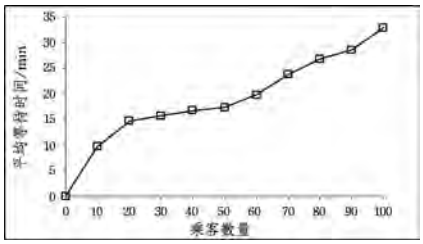


图 11　平均等待时间随乘客数量的变化趋势

图 12 给出了不同时间段对平均等待时间的影响。在工作日，平均等待时间有两个高峰值(上午 8 点左右和下午 7 点左右)，这两个值分别对应了上下班时间。由图可知，在高峰时间段，乘坐出租车的平均等待时间(请参见文献[12])比乘坐公交车的时间更长，但是其他大部分时间中乘坐出租车的等待时间更短。在周末，乘坐公交车和出租车在上午 10 点到下午 5 点这段时间的平均等待时间要长于其他时间段。这是因为，在烟台人们周末一般都选择在上午 10 点到下午 5 点之间出行，而晚上很少有活动。

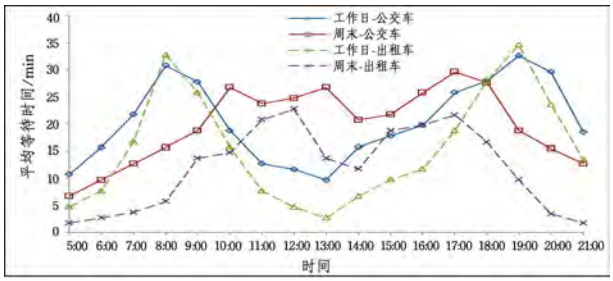


图 12　平均等待时间随时间的变化趋势

乘客比率定义为某一时刻的当前乘客数量与当天的乘客总数的比值。由图 13 可以看出，在工作日，人们的出行时间一般集中在上午 5 点到 9 点和下午 1 点到 5 点之间；而在周末，人们的活动时间通常集中在上午 10 点到下午 5 点之间。

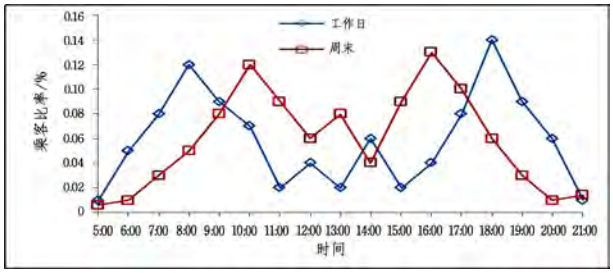


图 13　工作日和周末不同时间点的乘客比率

图 14 给出了 4 种模型(时变泊松模型、加权时变泊松模型、综合自回归移动平均模型、基于滑动窗口的整合框架)和 APM 的准确率随乘客数量的变化趋势。

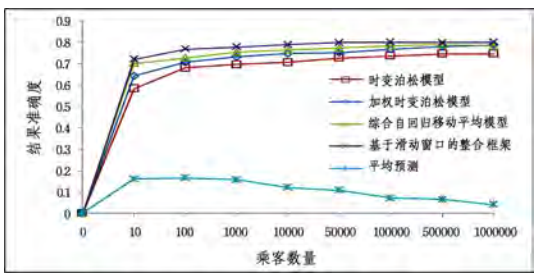


图 14　乘客数量对模型准确率的影响

当乘客数量趋于 0 时，所有模型的准确率也趋于 0，因为此时几乎没有结果产生。然而，当乘客数量稍微增大时，4 种模型的准确率都迅速增大。之后，随着乘客数量的增多，4 种模型的准确率几乎不变，即乘客数量对 4 种模型的准确率影响很少。而随着乘客数量的增多，APM 的准确率却缓慢下降。

图 15 显示的是聚合周期 P 对 4 种模型和 APM 准确率的影响。随着 P 的增大，APM 的准确率缓慢下降。然而，当 P 小于 30 min 时，4 种模型的准确率不受聚合周期的影响；当 P 大于 30 min 时，随着 P 的增大，4 种模型的准确率迅速下降。因此，本文将聚合周期设定为 30 min(即每 30 min 进行会做一次新的预测)。

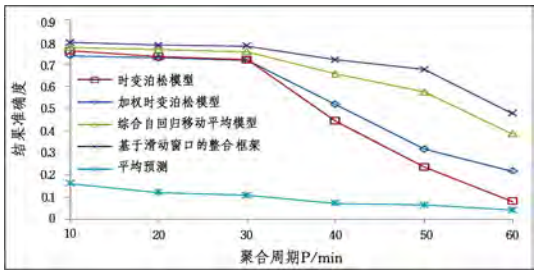


图 15 聚合周期对模型准确率的影响

我们在图 16 中测试了不同半径 W 对结果准确率的影响。当 W 趋于 0 时，所有模型的准确率都较低，因为此时我们无法预测即将到达的乘客数量。当 W 小于 100 m 时，随着 W 的增大，4 种模型的准确率增加，而参数 W 对 APM 的准确率影响很小。当 W 大于 100 时，结果的准确率趋于不变。因此，本文将半径设定为 100 m。参数 P 和 W 是根据公交站点上乘客的平均等待时间(大约 21 min)设定的。

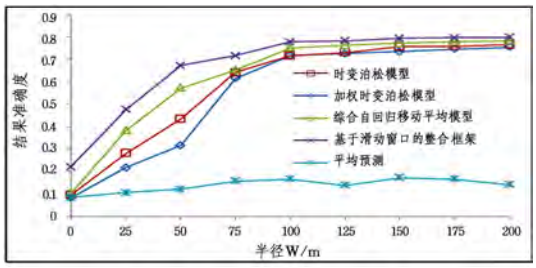


图 16 半径范围对模型准确率的影响

表 3 列出了当 $\alpha=0.5$ 时不同时间段内各模型的准确率。

表 3 模型的准确率($\alpha=0.5$)

(单位：%)

模型	时间段			
	上午 5 点到上午 9 点	上午 9 点到下午 1 点	下午 1 点到下午 5 点	下午 5 点到下午 9 点
时变泊松模型	75.83	71.69	73.58	74.13
加权时变泊松模型	76.63	73.58	74.39	74.86
综合自回归移动平均模型	78.35	73.79	75.99	76.28
基于滑动窗口的整合框架	79.37	76.62	78.06	78.33

由表可见，在上午 5 点到上午 9 点和下午 5 点到下午 9 点两个时间段内，模型的准确率高于上午 9 点到下午 1 点和下午 1 点到下午 5 点两个时间段。这是因上午 5 点到上午 9 点和下午 5 点到下午 9 点分别是上班和下班的时间，乘客人数相对固定。而上午 9 点到下午 1 点和下午 1 点到下午 5 点两个时间段内乘客自由活动的概率更大。

参数 α 从 0 到 1 变化时，我们测试了每种模型的准确率。由图 17 可以看出，参数 α 对 APM 的准确率影响很小。当 α 趋于 0 或者 1 时，4 种模型的准确率很低。其他情况下，结果的准确率趋于稳定。当结果趋于稳定时，只有加权时变泊松模型和整合框架受 α 变化的影响。基于滑动窗口的整合框架的准确率最高，接近 78%。

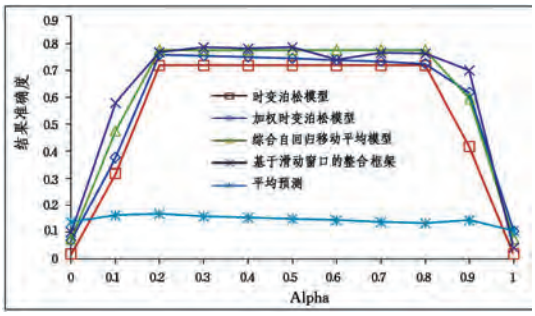


图 17 参数 α 对模型准确率的影响

我们设计了一个用户投票界面用来搜集用户的满意度。统计数据包含 2000 多位用户半年的投票数据，其中包括 1000 多名学生，以及政府工作人员、IT 工作者和其他人员若干。用户每完成一次查询，屏幕上就会显示 4 个选项：非常满意、满意、不满意和非常不满意。满意度(不满意度)是通过统计相应选项用户的数目除以总用户数目得到的。图 18 给出了乘客需求预测的满意度，几乎每位用户都对基于滑动窗口的整合框架的结果满意，其他 3 种模型以及 APM 的满意度都较低，其中 APM 的满意度最低。

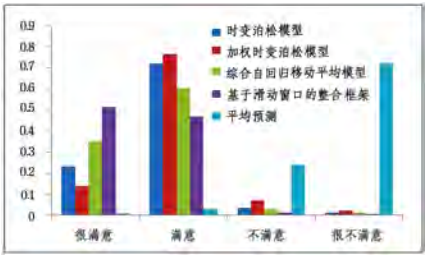


图 18 不同模型的满意度对比

结束语 本文提出了一个推荐系统用来预测公交服务中的乘客需求量，该系统中使用了 3 种预测模型。有效的乘客需求预测将成为公交网络中提供高级服务的一个重要的新特征，并且对其他基于位置的服务应用(Location Based Services，LBS)是非常有用的。有效的乘客需求预测可以帮助公交公司决定合理的公交发车时间间隔，减少乘客的等车时间，从而减少甚至避免交通堵塞。

参 考 文 献

[1] DAGANZO C F. A Headway-based Approach to Eliminate Bus Bunching: Systematic Analysis and Comparisons[J]. Transportation Research Part B Methodological, 2009, 43(10): 913-921.

[2] YAN S Y, CHI C J, TANG C H. Inter-city Bus Routing and Timetable Setting under Stochastic Demands[J]. Transportation Research Part A Policy & Practice, 2006, 40(7): 572-586.

[3] CHANG H，PARK D，LEE S，et al. Dynamic Multi-interval Bus Travel Time Prediction using Bus Transit Data[J]. Transportmetrica，2010，6(1)：19-38.

[4] YU B，LAM W，TAM M L. Bus Arrival Time Prediction at Bus Stop with Multiple Routes[J]. Transportation Research Part C Emerging Technology，2011，19(6)：1157-1170.

[5] DORNBUSH S，JOSHI A. StreetSmart Traffic：Discovering and Disseminating Automobile Congestion Using VANET's[C]// Proceedings of the 65th Vehicular Technology Conference，VTC2007-Spring. IEEE，2007：11-15.

[6] SCHUNEMANN B，WEDEL J，RADUSCH I. V2X-Based Traffic Congestion Recognition and Avoidance[J]. Tamkang Journal of Science and Engineering，2010，13(1)：63-70.

[7] LAKAS A，CHAQFEH M. A Novel Method for Reducing Road Traffic Congestion using Vehicular Communication[C]// Proceedings of the 6th International Wireless Communications and Mobile Computing Conference. IWCMC，ACM，2010：16-20.

[8] FERREIRA M，FERNANDES R，CONCEICAO H，et al. Self-organized Traffic Control[C]// Proceedings of the Annual International Conference on Mobile Computing and Networking. MOBICOM，ACM，2010：85-90.

[9] HOLT C. Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages[J]. International Journal of Forecasting，2004，20(1)：5-10.

[10] MAKRIDAKIS S，HIBON M. The M3-Competition：Results，Conclusions and Implications[J]. International Journal of Forecasting，2000，16(4)：451-476.

[11] ZHOU C，MENG X，CHEN Y. Out-of-order Durable Event Processing in Integrated Wireless Networks[J]. Pervasive and Mobile Computing，2011，7(5)：595-610.

[12] CHANG H，TAI Y，CHEN H，et al. iTaxi：Context-Aware Taxi Demand Hotspots Prediction Using Ontology and Data Mining Approaches[C]// Proceedings of the 13th Conference on Artificial Intelligence and Applications (TAAI). 2008.

（上接第 507 页）

表 4　使用扩维技术之后的训练结果

阶段	循环	时间	准确率/%
5	250	34.88	92.97
7	400	53.87	96.09
8	450	61.05	98.44
9	500	75.04	100.00
10	550	81.84	100.00
11	600	90.84	100.00
12	650	100.05	100.00
13	700	106.15	100.00
13	750	115.66	100.00
14	800	121.62	100.00
15	850	135.64	100.00
15	870	142.20	100.00

2　研究背景

卷积网络是当前深度学习应用得最广泛的模型之一，一般用于与图像有关的领域，例如计算机视觉中的物体识别、语义识别、基于向量化的词语。

在前期手背纹理识别生理年龄的研究中注意到，图像应用中复杂背景的影响和每个人的一些生理数据特征很像，例如性别相同的人却有不同的心电图模式等，因此将其应用于不同时变特性和不同维度的生理信号的多元时间序列处理，分类的目标就是中医中的“脉象”。

脉象是中医中重要的“诊病”参考，但对大多数“不熟悉”中医的人来说，脉象非常神秘。明代李士材在《诊家正眼》中在李时珍的基础上增加了疾脉，形成 28 脉象并为后世沿用。

显然，脉象的分类是典型的机器学习问题，将手指感知的脉搏跳动转化为不同的脉像描述，我们试图将越来越普遍的生理信号捕捉设备与脉象联系起来，改变目前“脉象仪”固定感知且无法迁移的模式，同时也为更好地“量化”传统医学的深厚经验，并为更多人所用。

结束语　本文提出了一种针对时变特征差异较大的向量时间序列的卷积网络分析方法，同时利用多分辨率分析方法和统计特征分析来实现“扩维”，有效降低了训练强度。该研究应用于“基于生理数据”分析脉象特征，通过测试样本进行分类和对比，准确率可以超过 90%。

致谢　感谢数据灯塔团队的讨论和支持。

参 考 文 献

[1] 张宁. 深度学习改变保险精算定价模式[J]. 计算机科学，2017，44(3)：1-2.

[2] 刘琮，许维胜，吴启迪. 时空域深度卷积神经网络及其在行为识别上的应用[J]. 计算机科学，2015，42(7)：245-249.

[3] ALTWAIJRY H，TRULLS E，HAYS J，et al. Learning to match aerial images with deep attentive architectures[C]// Computer Vision and Pattern Recognition. IEEE，2016：3539-3547.

[4] XIA L，et al. Selected by input：Energy efficient structure for rrambased convolutional neural network[C]// DAC. 2016

[5] HILTON G E，SALAKHUTDINOV R R. Reducing the Dimensionality of Data with Neural Network[J]. Science，2006，313：504-507.