文章编号：2095-2163(2020)03-0155-05 中图分类号：U231 文献标志码：A

基于 PCA-LSTM 模型的城市轨道交通短时客流预测

石敏莲[1,2]，刘志钢[1]，胡 华[1]，汪 景[1]
（1 上海工程技术大学 城市轨道交通学院，上海 201620；2 美国威斯康辛大学 土木工程学院，Wisconsin 53211）

摘 要：城市轨道交通的进出站客流量具有较大的不确定性和复杂性，尤其是短期客流预测，一直是地铁客流预测中的一个研究热点和难点。AFC 设备能准确读取刷卡数据，实现历史和实时进出站客流量的有效统计。为提高进出站客流预测精度，本文以杭州地铁西兴站为例，利用主成分分析法（PCA）对通过 AFC 设备采集的历史进出站客流数据进行特征提取，然后通过处理后的数据建立长短期记忆网络（LSTM）短期客流预测模型。仿真结果表明该方法在城市轨道交通进出站客流预测中有较好的表现，满足短期客流预测的要求，能够为地铁的运营管理提供一定的指导作用。

关键词：短期预测；客流；PCA；LSTM

Short-term passenger flow forecast of urban rail transit based on PCA-LSTM model

SHI Minlian[1,2], LIU Zhigang[1], HU Hua[1], WANG Jing[1]
(1 School of Urban Rail Transportation, Shanghai University of Engineering Science, Shanghai 201620, China;
2 School of Civil Engineering, University Wisconsin Milwaukee, Wisconsin 53211, USA)

【Abstract】 The passenger flow of urban rail transit in and out of the station is of great uncertainty and complexity, so it's hard to forecast the volume of it in short-term. AFC equipment can accurately read card data, then realize the history and real-time statistics of passenger flow in and out of the station. In order to improve the prediction accuracy of forecast of number of people arriving or leaving the station, this paper takes Hangzhou Xixing Station as an example. Firstly, the principal component analysis (PCA) method is used to extract the characteristics of the historical passenger flow data collected by AFC equipment, and then establishes the short-term and long-term memory network (LSTM) short-term passenger flow prediction model through the processed data. The simulation results show that this method has a good performance in the passenger flow prediction in and out of the station of urban rail transit, meets the requirements of short-term passenger flow prediction, which can provide some guidance for the operation and management of the subway.

【Key words】 short-term forecast; passenger flow; PCA; LSTM

0 引 言

随着社会经济的飞速发展，人们的生活节奏加快，出行频率也大幅度增加，同时对出行效率和舒适度的要求也越来越高。对于城市轨道交通而言，客流量是运营的主要依据，也是构建智慧交通的重要基础。日常列车排班计划的制定、大客流的预防等均要求对未来客流量进行预测。

对于短期客流预测，主要可分为线性和非线性两类。其中，线性预测常用方法有卡尔曼滤波、时间序列预测等；非线性预测常用方法主要包括灰色理论、神经网络、支持向量机等。近年来，国内外许多专家学者对这类客流预测进行了大量的研究。王奕等人[1]根据周期时变特点在灰色预测模型的基础上改进了马尔科夫算法。杨军[2]将小波分析与支持向量机结合提出了短期客流预测方法。程浩等人[3]利用 BP 神经网络对短期客流进行预测。侯晨煜等人[4]在神经网络算法的基础上，结合卡尔曼滤波，提出了一种新型有效的地铁客流短时预测算法。Han 等人[5]提出了一种新的基于深度学习的方法 STG-CNN (spatial-temporal graph convolutional neural networks for metro)，对城市每个地铁站的进站流量和出站流量进行了综合预测。Sun 等人[6]提出了一种新的混合模型小波-支持向量机，结合了小波与支持向量机模型的互补优势，同时克服了其

基金项目：十三五国家重点研发计划子课题(2017YFC0804900)；上海市科委地方院校能力建设项目(19030501400)；国家自然科学基金(71601110)；同济大学道路与交通工程教育部重点实验室开放基金(K201902)。

作者简介：石敏莲(1990-)，女，双学位硕士，主要研究方向：客流预测、列车运行图优化；刘志钢(1974-)，男，博士，教授，上海工程技术大学城市轨道交通学院院长，主要研究方向：城市轨道交通运营管理优化及安全技术。

通讯作者：刘志钢 Email：zzxcsml@126.com

收稿日期：2019-10-15

各自的不足。但是,较少有学者把预测站点与其他站点的客流相关性放入预测模型中进行综合考虑。

本文以杭州地铁西兴站为例,考虑到站点之间客流的空间和时序相关性,利用主成分分析法(PCA)对通过 AFC 设备采集的历史进出站客流数据进行特征提取,然后通过处理后的数据建立长短期记忆网络(LSTM)短期客流预测模型并进行模型有效性验证。

1 短期客流预测

对城市轨道交通短期客流预测的研究能为突发性大客流的预防和列车调度的优化提供有力的参考。现有的短期客流预测一般以 15~60 min 为时间粒度,指根据历史客流和实时客流等数据,利用客流预测模型,计算得到预测对象在 15 min 后的客流情况,若该数值超过行业规范或运营公司所给出的安全范围,则相关运营部门和工作人员应按照相应的安全预案立刻开展行动,如通过广播播报、入口限流等措施来保障车站以及站台人流密度在安全范围内,预防踩踏等危及乘客人身安全事件的发生,确保乘客的安全和列车的正常运营。而以 60 min 为时间粒度进行客流预测,能够为列车调度的优化提供依据,通过调整列车运行计划提高运输效率或节约运营成本。列车运行计划的调整,一般情况下,并不能在 15 min 内即刻完成。例如,根据客流需求的意外增长,某线路产生了加开一班列车的需求,调度部门需先结合原有排班计划调整列车运行图,再通过部门审批、车辆段对上线列车进行准备工作,还需通知司机等相关执行人员等,整个过程需要 30 min~1 h。因此,以 1 h 为长度对车站进出站客流进行预测,对列车运行实时优化具有十分重要的意义。

2 PCA-LSTM 预测模型

2.1 PCA 特征提取

在实验和研究的过程中,经常会遇到这样的情况,即对同一研究对象存在大量影响因素。越全面的数据确实能为实验目的提供越丰富的信息,但是同时也会提高模型的计算和训练时间。而且,许多变量之间可能存在较大的相关性或相似性。因此,盲目地增加变量可能会极大地加长运算时间,但是对研究目的产生的帮助甚微,而盲目地减少变量可能会损失重要的信息,影响结论的准确性。

PCA 法就是一种对多维数据进行降维的数据预处理方法[7]。通过计算分析各维度数据之间的相关性,PCA 法能去除多维数据中一部分不重要的特征,保留相对重要的那部分,从而使得数据更易于使用,提升计算速度。PCA 法主要思想是将 n 维数据映射到 k 维上,且这 k 维的特征向量相互正交。特征向量的选取标准是取特征值最大的 k 个特征所对应的特征向量,目的是使得这 k 为数据尽量多的保留原数据的特征,减少信息损失。新构造的维度对原维度数据信息的反映一般通过方程贡献率来衡量。一般会选取累计贡献率为 80%~95%的 k 维数据作为降维后数据。

在城市轨道交通客流预测研究中,历史客流数据是进行客流预测的最主要、也是最直接的依据。在对某一站点进行客流预测时,一般该站点的历史进出站客流数据作为主要因素,再结合其他因素,作为预测模型的输入。其实,除了预测站点自身的历史客流数据外,同一线网中的其他的车站的客流进出量也能为该车站的客流预测提供很好的参考。例如 A 站点在某时间段内进站客流的增加,有一定的可能性使得 B 站点在下一时间段的出站客流增加。再如,首发站点 A 站在这一时间段内进站客流增加较大,则其后续站点在之后的短时间内进站客流增加的概率较大。

然而,对大多数城市来说,整个地铁线网的数据量过于庞大,就上海地铁来说,一共有 16 条线路,共有 415 座车站(含 2 座磁悬浮线车站)。即使就单一一天地铁线路来讲,其站点数量也不少,例如杭州地铁 1 号线,一共有 34 个车站。若使用所有站点的历史进出站数据,会极大地提高计算复杂性和计算时间,导致计算机无法在有限时间内给出相应的预测结果。因此,为提高模型训练速度并降低计算复杂性,本文采用主成分分析方法对线路上的进出站客流数据进行降维。

选取杭州轨道交通一号线在 2018 年 12 月 20 日~ 2019 年 5 月 9 日期间沿线各站点运营时段每小时(5:00~7:00 时段数据合并为一个数据)进出站客流量作为实验数据。把全天运营时间按顺序划分为 20 个时段,见表 1,每时段采集一次线路上各站点的进出站客流数据。一号线一共有 34 个车站,每个车站采集各时段进站客流和出站客流两组数据,全线共有 68 组数据。同时,数据采集时段与各车站客流之间的关系非常密切,故将运营时段进行编号后放入影响因素集中,详见表 1。此时数据集为 69 维。

选定某站点进站或出站客流作为预测对象,文中随机选择了西兴站出站客流作为预测目标,因此先从 69 维数据集中抽取出西兴站的出站客流数据

以备后用，将剩余的 68 维数据通过 PCA 法进行降维，得到新的变量。根据方差贡献率和累计贡献率，从高到低，选择主成分，将原来的 68 个变量压缩成 4 个主成分，保留了原始数据约 90%的信息，得到的主成分方差贡献率和累计贡献率见表 2。

表 1　运营时段划分表

Tab. 1　Table of operating hours

时段	编号	时段	编号
5:00-7:00	0	16:00-17:00	10
7:00-8:00	1	17:00-18:00	11
8:00-9:00	2	18:00-19:00	12
9:00-10:00	3	19:00-20:00	13
10:00-11:00	4	20:00-21:00	14
11:00-12:00	5	21:00-22:00	15
12:00-13:00	6	22:00-23:00	16
13:00-14:00	7	23:00-24:00	17
14:00-15:00	8	0:00-1:00	18
15:00-16:00	9	1:00-2:00	19

表 2　方差及方差贡献率

Tab. 2　Variance and variance contribution rate

排名	方差	方差贡献率	累计贡献率
1	46 794 179.996	0.616	0.616
2	12 413 647.377	0.163	0.779
3	6 832 409.073	0.090	0.869
4	2 389 622.804	0.031	0.900

将西兴站出站数据与降维得到的 4 个主成分数据合并，得到维度为 5 的变量数据作为预测模型的输入。

2.2　LSTM 网络

LSTM 网络是循环神经网络的一种，是为了解决普通循环神经网络（RNN）所存在的梯度易消失和长期记忆被遗忘的缺点而提出的[8-10]。RNN 网络主要由重复的神经网络模块进行链式组合而成，每个模块有 2 个输入数据和 2 个输出数据。LSTM 网络在 RNN 网络的基础上增加了一个输入和一个输出，内部结构也更为复杂精细。增加的这一路输入和输出称为细胞状态，是 LSTM 实现状态记忆和遗忘的主要结构，上面的信息与当前状态的输入信息仅有 2 次线性交互，使得细胞状态较容易保持稳定，达到长期记忆的目的。

LTSM 网络的细胞结构如图1所示，主要由输入门、遗忘门和输出门组成。细胞内信息的处理需要经过以下 4 个步骤：

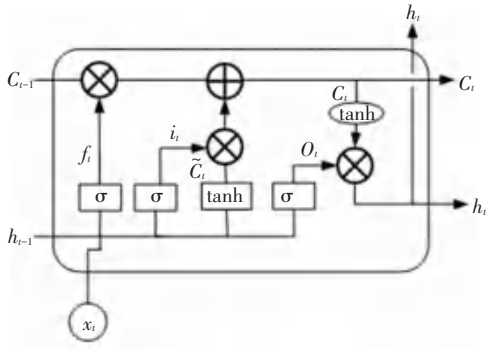


图 1　LSTM 网络细胞结构

Fig. 1　Cell structure of LSTM

（1）遗忘门，输入为上一层的输出 h_{t-1} 和当前层的状态 x_t。遗忘门在读取数据后利用 sigmoid 函数将信息转化为 0~1 之间的一个数值，来表示对这两个输入信息的遗忘程度，其中 0 表示完全遗忘，1 表示完全保留。此时对应的数学公式可表示为：

$$f_t = \sigma(\boldsymbol{w}_f \cdot [k_{t-1}, x_t] + \boldsymbol{b}_f), \tag{1}$$

（2）输入门，这一步决定了让多少新的信息加入到当前细胞状态中来，主要通过 2 个函数实现。首先利用 sigmoid 函数决定需要更新的信息 i_t，再用 tanh 函数产生 $\tilde{C}_t$，作为备用信息。此时对应的数学公式可分别表示为：

$$i_t = \sigma(\boldsymbol{w}_i \cdot [h_{t-1}, x_t] + \boldsymbol{b}_i), \tag{2}$$

$$\tilde{C}_t = \tanh(\boldsymbol{w}_c \cdot [h_{t-1}, x_t] + \boldsymbol{b}_c), \tag{3}$$

（3）状态更新，这一步结合了数据分别经过遗忘门和输入门得出的遗忘状态值 f_t 和更新状态值 i_t 来进行信息的遗忘和更新，从而获得新的细胞状态 C_t。此时对应的数学公式可表示为：

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t, \tag{4}$$

（4）输入门，用于确定该层细胞需要输出的信息。首先，用 sigmoid 函数来确定初始输出 O_t，表示需要保留并输出的信息。接着，把当前细胞状态 C_t 通过 tanh 函数进行处理，把所得结果与初始输出信息 O_t 相乘，得到最终输出 h_t。此时对应的数学公式可表示为：

$$O_t = \sigma(\boldsymbol{w}_0[h_{t-1}, x_t]) + \boldsymbol{b}_0, \tag{5}$$

$$h_t = O_t \cdot \tanh(C_t). \tag{6}$$

式（1）~（6）中，$\boldsymbol{w}_f, \boldsymbol{w}_i, \boldsymbol{w}_c, \boldsymbol{w}_0$ 为各自对应的权重矩阵，$\boldsymbol{b}_f, \boldsymbol{b}_i, \boldsymbol{b}_c, \boldsymbol{b}_0$ 为各自对应的偏置向量。

LSTM 网络模型和 RNN 网络模型相似,都由循环的网络组成,将其按时序展开以后可看作链式结构,信息在各层细胞内经过选择、遗忘和更新后,在细胞之间传递。LSTM 网络细胞的时序展开的释义解析如图 2 所示。

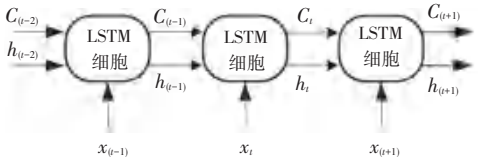


图 2 LSTM 网络细胞的时序展开图

Fig. 2 Sequence diagram of LSTM network cells

2.3 运用 LSTM 网络进行预测

2.3.1 参数配置

建立该 LSTM 网络预测模型需要确定一些超参数,包括输入层的维数、隐藏层的层数与维数、时间步长以及输出层的维数。

本实验以西兴站出站客流量为预测对象,将其历史数据与 PCA 降维得到的 4 维变量数据一起作为 LSTM 网络的输入,该 LSTM 网络输入层维数为 5。预测目标为下一小时出站客流量,确定时间步长为 1,输出层维数为 1。经过多次尝试,确定隐藏层为 2 层,第一层中神经元数量为 50 个,第二层中神经元数量为 30 个。选定 Adam 优化器作为该 LSTM 网络的优化算法。

首先,将输入数据集按照 8:2 的比列划分为训练集和测试集,训练集包含了 2 223 组数据,测试集包含了 557 组数据。使用基于 TensorFlow 后端的 Kersa 框架进行编程测试,并使用训练集对该神经网络模型进行训练,训练次数设定为 200 次。之后利用测试集对模型预测结果进行验证和评估。评估的方式采用均方根误差(root mean square error, *RMSE*),其计算公式为:

$$\varepsilon_{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}. \tag{7}$$

其中,y_i 和 $\hat{y}_i(i = 1,2,3,4,\cdots,m)$ 分别表示第 i 个数据取样点的实际值和预测值,m 表示数据长度。

2.3.2 预测结果

LSTM 网络的训练结果如图 3 所示。观察图 3 发现,在前 25 次训练中,损失函数就下降到了 0.003 以下,收敛速度快,在训练后期,损失函数基本在 8.5×10^{-4}左右波动,模型训练时间为 117 s,表明模型训练效率高,而且训练效果比较理想。

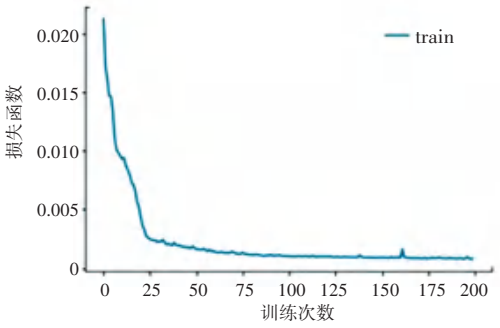
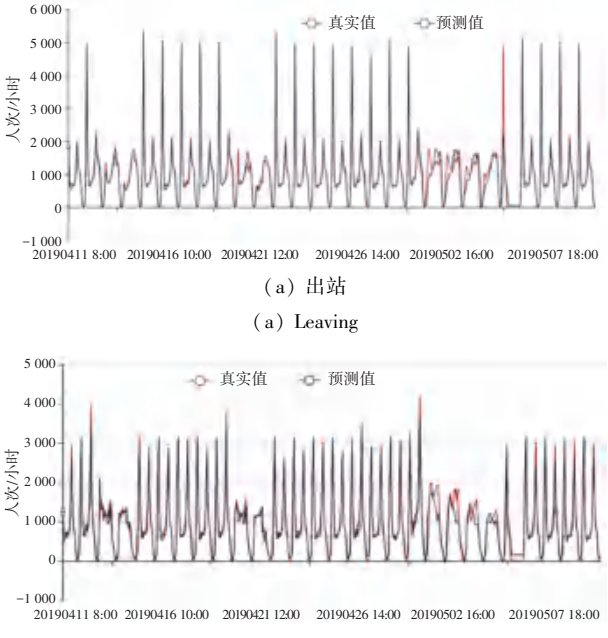


图 3 损失函数图

Fig. 3 Loss function graph

将测试集数据输入训练好的预测模型,获得预测结果如图 4 所示。由图 4 可知,PCA-LSTM 模型预测精度较高,预测误差 ε_{RMSE} 为 198.32。实验结果验证了该模型具有良好可行性和适用性。



(a) 出站

(a) Leaving

(b) 进站

(b) Arriving

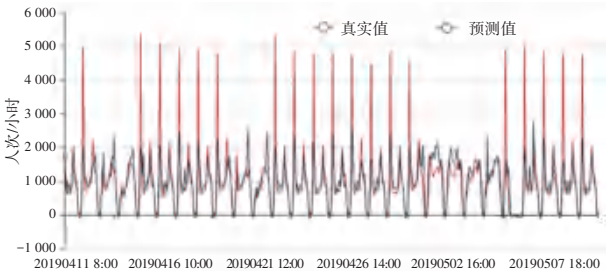
图 4 西兴站进出站实际客流数据与预测数据对比图

Fig. 4 Comparison chart of actual and forecast arriving and leaving passenger flow data of Xixing station

另外,使用同样的方法,将西兴站下一小时出站客流作为预测目标,重新训练模型,所得预测结果如图 5 所示。研究推得其 *RMSE* 为 239.23。说明PCA-LSTM 模型对进站客流和出站客流的预测均适用。

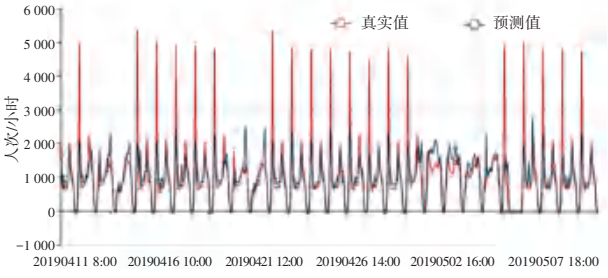
图 5 是以西兴站出站和进站客流数据为输入数据的单一 LSTM 网络模型所得的预测结果,训练时间分别为 59.49 s 和 65.93 s,其测试集的 *RMSE* 分别为 606.33 和 386.78。两者的 *RMSE* 分别是 PCA-LSTM 模型所得结果的 3.06 倍和 1.61 倍,见表 3。说明 PCA-LSTM 的预测精度与单一 LSTM 模型相比,有显著提

升，尤其是对出站客流的预测方面，优势明显。



（a）出站

（a）Leaving



（b）进站

（b）Arriving

图 5　实际值与 LSTM 模型预测结果对比图

Fig. 6　Comparison diagram of actual value and prediction results of LSTM model

表 3　均方根误差对比表

Tab. 3　*RMSE* comparison table

预测变量	*RMSE*		比例
	PCA-LSTM	LSTM	
西兴站出站客流	198.32	606.33	3.06
西兴站进站客流	239.23	386.78	1.61

3　结束语

本文从同一地铁线路上车站客流之间存在相关性这一角度出发，设计了基于 PCA-LSTM 的城市轨道交通短时客流预测模型，采用了杭州地铁一号线 139 天的进出站客流数据进行预测实验。结果表明，该模型在对站点下一小时进站客流量和出站客流量的预测方面具有较好的表现，能够为地铁运营部门在实际的列车运行优化和调度方面提供可靠的参考。该方法同样适用于以 15 min、30 min 等其他时间粒度的短期客流预测。未来的研究工作可以考虑把天气以及是否为工作日等其他因素加入到影响因素集中，从而进一步提高模型的预测精度。

参考文献

[1] 王奕，徐瑞华. 基于周期时变特点的城市轨道交通短期客流预测研究[J]. 城市轨道交通研究，2010，13(1)：46.

[2] 杨军. 地铁客流短期预测及客流疏散模拟研究 [D]. 北京：北京交通大学，2014.

[3] 程浩，徐昕. 基于 BP 神经网络的轨道客流短期预测 [J]. 电子技术与软件工程，2016(22)：15.

[4] 侯晨煜，孙晖，周艺芳，等. 基于神经网络的地铁短时客流预测服务 [J]. 小型微型计算机系统，2019，40(1)：226.

[5] HAN Yong，WANG Shukang，REN Yibin，et al. Predicting station-level short-term passenger flow in a citywide metro network using spatiotemporal graph Convolutional Neural Networks [J]. ISPRS International Journal of Geo-Information，2019，8(6)：243.

[6] SUN Yuxing，LENG Biao，GUAN Wei. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system [J]. Neurocomputing，2015，166：109.

[7] 白亚男. 基于大数据的实时交通流预测方法研究 [D]. 广州：广东工业大学，2018.

[8] 晏臻，于重重，韩璐，等. 基于 CNN+LSTM 的短时交通流量预测方法 [J]. 计算机工程与设计，2019，40(9)：2620.

[9] 张铭坤，王昕. 基于 GRU-RNN 模型的城市主干道交通时间预测 [J]. 北京信息科技大学学报(自然科学版)，2019，34(4)：30.

[10] 崔洪涛，陈晓旭，杨超，等. 基于深度长短期记忆网络的地铁进站客流预测 [J]. 城市轨道交通研究，2019(9)：41.

(上接第 154 页)

块链技术的服装产业协同制造溯源系统。但是目前区块链技术溯源系统数据吞吐量不高，使得区块链技术溯源平台推广应用有非常大的障碍。后续工作需要进一步研究区块链技术，并优化溯源系统性能，才能将区块链技术溯源平台进行广泛商业应用。

参考文献

[1] 鲁维维. 区块链技术在供应链管理中的应用研究[J]. 当代经济，2017(29)：98.

[2] 孙筱，赵洪珊. 服装产品开发流程管理解析[J]. 山东纺织经济，2013(11)：13.

[3] 杨年生. 纺织服装业如何跨界区块链[J]. 纺织科学研究，2019(4)：42.

[4] JAOUDE J A，SAADE R G. Blockchain applications-usage in different domains[J]. IEEE Access，2019，7：45360.

[5] NAKAMOTO S. Bitcoin：A peer-to-peer electronic cash system [EB/OL].[2019-05-28]. https://bitcoin.org/bitcoin.pdf.

[6] FREISE M，SEURING S. Social and environmental risk management in supply chains：A survey in the clothing industry [J]. Logistics Research，2015，8(1)：1.

[7] Al-JAROODI J，MOHAMED N. Blockchain in industries：A survey[J]. IEEE Access，2019，7：36500.

[8] SZABON. Smart contracts[EB/OL].[1994]. http://szabo.best.vwh.net/smart.contracts.html.

[9] MERKLE R C. Protocols for public key cryptosystems[C]// 1980 IEEE Symposium on Security and Privacy. Oakland，CA，USA：IEEE，1980：122.

[10] CARTER J L，WEGMAN M N. Universal classes of hash functions[J]. Journal of Computer & System Sciences，1979，18(2)：143.

[11] DIFFIE W，HELLMAN M. New directions in cryptography[J]. IEEE Transactions on Information Theory，1976，22(6)：644.