计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600001

基于数据挖掘的指定航班计划延误预测方法

张成伟　罗凤娥　代　毅
中国民用航空飞行学院空中交通管理学院　四川 广汉 618300

摘　要　针对现有航班延误预测方法较少从指定航班计划延误预测角度进行分析,提出一种研究离港航班计划中指定某航班计划发生延误情况的预测方法。首先,分析大量航班历史运行数据,挖掘数据内在特征。其次,通过建立航班数据的贝叶斯网络分析模型,得到不同条件下航班延误情况的概率分布;以动态贝叶斯网络(Dynamic Bayesian Networks,DBN)推理为主要建模方法,研究了动态贝叶斯网络推理和仿真过程,提出了一种用于构建航班延误预测模型的新方法,建立了实际航班数据的隐马尔可夫(Hidden Markov Model,HMM)延误预测分析模型,利用隐马尔可夫模型中解码问题 Viterbi 算法实现了指定航班延误时间的预测。最后,以某航空公司全年航班运行数据为例进行实例仿真及验证,结果表明,该方法实现了航班延误预测对象的精确性。

关键词:指定航班计划;数据挖掘;贝叶斯网络;隐马尔可夫模型;延误预测
中图法分类号　F560

Prediction Method of Flight Delay in Designated Flight Plan Based on Data Mining

ZHANG Cheng-wei,LUO Feng-e and DAI Yi
College of Air Traffic Management,Civil Aviation Flight University of China,Guanghan,Sichuan 618300,China

Abstract　In view of the fact that the existing flight delay prediction methods are rarely analyzed from the perspective of the designated flight plan delay prediction,a prediction method to study the delay situation of a specified flight plan in the departure flight plan is proposed. First,analyzing the intrinsic characteristics of a large number of historical flight data mining data. Secondly,this research employs Dynamic Bayesian Network inference as the main modeling method to obtain the probability distribution under different conditions of flight delay. By studying the Dynamic Bayesian Network inference process and simulation,this paper presents a new method for the construction of the flight delay prediction model which is to establish Hidden Markov flight delay prediction model based on the real flight data. Using the Viterbi algorithm of Hidden Markov model decoding problem to predict the flight delay time. Finally,taking an airline's full-year flight operation data as an example for example simulation and verification,the results show that this method improves the accuracy of flight delay prediction objects.

Keywords　Designated flight schedule,Data mining,Bayesian networks,Hidden Markov model,Delay prediction

1　引言

航班延误问题是限制我国民航业快速发展的亟待解决的问题之一,随着新成立航空公司的不断涌现,航班延误情况直接影响到公司的品牌效应,同时也对航空公司的管理工作造成很大的麻烦。若想缓解航班延误问题,需从运行控制的角度首先针对航班计划执行情况进行预测。

在航空公司运行控制指挥中心(Airline Operation Control Center,AOC)制定的 3 天以内航班实际运行过程中,航班延误时常发生。签派员往往不能确定已延误的航班延误的时间有多长,进而对后续航班的保障能力有限,运行控制的效果受到一定的限制。

近年来,国内外诸多学者对航班延误预测方面的研究,已经取得一定的成果。在国外,2004 年 Fabdelghany 等[1]以有向无环图的形式来表示航班计划,基于最短路径算法对延误时间进行推理分析,但研究的缺点在于不能预测航班的延误状态;2005 年 Tu 等[2]重点选取离港航班作为研究的主要对象,以人工智能相关算法为主要预测建模方法,模型检验采用混合分布残差估计,解决了混合模型中局部最优值的问题,该研究模型拟合度高,且具有较高的预测能力。2007 年 Xu 等[3]研究了航班延误各个因素之间的关联性,通过静态贝叶斯网络分析模型的建立,运用线性和非线性回归方法仿真航班延误影响因素所带来的影响,该研究同时指出研究影响航班延误的多种因素的重要意义。2020 年,Arora 等[4]使用服务恢复双偏差框架来描述航班延误到达的可能性,利用多项式 logistic 回归方法建立航班到达延误时间预测模型,分析了

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:民航局安全能力建设项目(OMSA1805);中央高校教育教学改革专项(E20180302);中国民用航空飞行学院青年基金项目(XM4043);航空运行控制技术研究所(JG201935)
This work was supported by the CAAC Safety Capacity Project(OMSA1805),Central University Education and Teaching Reform(E20180302),CAFUC Youth Fund Project(XM4043) and Aviation Operation Control Technology Institute(JG201935).
通信作者:张成伟(dasen.lin.ok@163.com)

美国不同航空公司运行控制的特点。在国内，2008 年，Yao 等[5]实现了对不正常航班进行预警的效果，并将预警级别分为 5 个等级，选取延误人员数量、延误率以及可用资源等指标建立预警评价指标体系，其对不正常航班管理等相关方面也做了大量深入研究。2009 年，Xu 等[6]研究了基于贝叶斯网络模型的航班延误波及分析，根据得到的延误情况概率分布进行相应预测，并对各因素之间的相互影响进行了研究。2010 年，Lv 等[7]从航班延误导致的结果入手，从机场运行角度出发提出预警指标体系的构建。2018 年，Wu 等[8]通过数据融合，基于双通道卷积神经网络建立了航班延误预测模型，并选择 Padding 技术进行了模型优化。2019 年，Wu 等[9]基于深度 SE-DenseNetDE，融合航班信息及气象信息，构建航班延误预测分析模型。2020 年，Zhang 等[10]从空中交通网络流系统角度界定大面积航班延误，计算基本再生数预测大面积航班延误能否发生。

上述航班延误预测方面的研究虽已取得一定的成果，但多是从航班周期运行角度出发对延误进行预测，未研究针对某一指定航班的延误预测方法。鉴于此，拟针对某一指定航班计划进行延误预测分析，预测对象为某架指定飞机在已知离港机场所要执行的航班计划。以数据挖掘技术中的贝叶斯网络为研究方法，建立实际航班数据的(BN)模型，挖掘延误要素之间的相互关联性。借鉴动态贝叶斯网络(DBN)仿真推理思想，利用隐马尔可夫解码问题 Viterbi 算法实现航班延误时间的预测，并在 MATLAB 环境下通过航班实际运行数据的仿真实验，来预测具体航班的延误可能性。

2 动态贝叶斯网络及隐马尔可夫模型

以往的数据挖掘以贝叶斯网络(Bayesian Networks，BN)为基础，通过网络结构与数据匹配找到变量间的关系，但这种方式往往耗时太多，不能用于实时性的推理进而影响航班实时计划延误预测的精确性。动态贝叶斯网络(Dynamic Bayesian Networks，DBN)基于静态贝叶斯网络模型又进一步进行了深入研究，形成了与时序数据相关的实时数据挖掘体系。

2.1 动态贝叶斯网络模型

动态贝叶斯网络的推理主要是依据大量观测数据，对其中的隐含变量的最大可能取值概率进行推测的过程。主要针对隐变量离散动态网络进行推理，构建基于 DBN 网络的航班延误预测分析模型，因此针对隐变量离散 DBN 给出推理过程[11]。

DBN 推理时可根据所建立 DBN 的性质对网络模型进行简化。隐马尔可夫模型(Hidden Markov Model，HMM)是所有离散 DBN 的基础，将复杂动态贝叶斯离散网络模型转化为简单 HMM 的方法，根据前向后向算法实现 DBN 的推理。

若记随机变量集为 $X=\{X_1,X_2,\cdots,X_n\}$，$Pa(X_i)$代表父节点集。在 t 时刻的X_i表示为$X_i[t]$。

$$P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i\mid Pa(X_i)) \tag{1}$$

DBN 模型引入含时间因素的随机过程。为了实现可行的建模，做如下假设：

(1)假设时间片内的时序变化是平稳的。

(2)假设该动态系统变化过程是具有马尔可夫性的，即满足式(2)。

$$P(X[t+1]\mid X[1],X[2],\cdots,X[t])=P(X[t+1]\mid X[t]) \tag{2}$$

(3)假设 DBN 的组成：一个先验网B_0，定义在初始状态 $X[1]$上的概率分布；一个转移网$B_{\rightarrow}$，定义在变量 $X[1]$ 与 $X[2]$上的转移概率 $P(X[t+1]\mid X[t])$。

先验网络B_0图和转移网络$B_{\rightarrow}$图如图 1 所示。因此，若给定一个 DBN 模型，则在 $X[0],X[1],\cdots,X[T]$上的联合概率分布为：

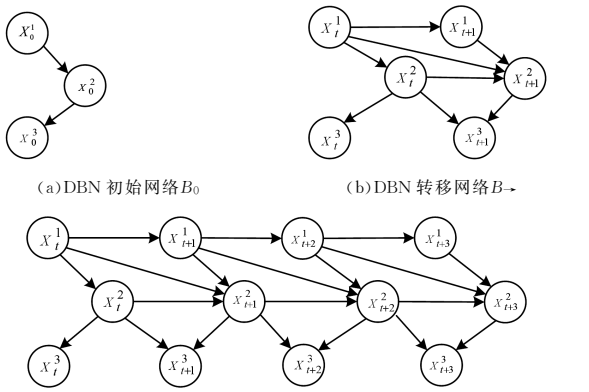
$$P(X[0],X[1],\cdots,X[T])=P_{B_0}(X[1])\prod_{t=1}^{T}P_{B_{\rightarrow}}(X[t+1]\mid X[t]) \tag{3}$$

若用 $P(X_t\mid X_{t-1})$表示某变量前一时刻状态在当前状态发生的概率，X_t^i 表示第 i 个变量 t 时刻的取值，$Pa(X_t^i)$表示其父节点，且当只有两个时间片段和 N 个变量时，有：

$$P(X_t\mid X_{t-1})=\prod_{i=1}^{N}P(X_t^i\mid Pa(X_t^i)) \tag{4}$$

同样 DBN 中任一节点的联合概率计算方法如式(5)所示：

$$P(X_{1:T}^{1:N})=\prod_{i=1}^{N}P_{B_0}(X_1^i\mid Pa(X_1^i))\times\prod_{t=2}^{T}\prod_{i=1}^{N}P_{B_{\rightarrow}}(X_t^i\mid Pa(X_t^i)) \tag{5}$$



(a)DBN 初始网络B_0　(b)DBN 转移网络$B_{\rightarrow}$

(c)按时间轴展开的 DBN

图 1 动态贝叶斯网络示意图

Fig. 1 Dynamic Bayesian Network diagram

根据大量航班运行历史数据中挖掘出航班延误影响因素之间的关联性来完成预测与预警，因此预测的推理计算式如式(6)所示：

计算 $P(X_{t+h}\mid y_{1:t})$，已知$y_{1:t}$的观察值，预测出 $t+h$ 时刻 X_{t+h}的值，即：

$$P(Y_{t+h}=h\mid y_{1:t})=\sum_{x}P(Y_{t+h}=y\mid X_{t+h}=x)P(X_{t+h}=x\mid y_{1:t}) \tag{6}$$

2.2 隐马尔可夫模型

隐马尔可夫模型是隐变量离散 DBN 的特例，任何一个离散 DBN 都可以转化为若干个标准 HMM 进行推理。图 2 所示为一个简单的 HMM 的图形模式，其中节点 y_1,y_2,y_3 表示观测节点，节点 x_1,x_2,x_3 表示隐含节点。

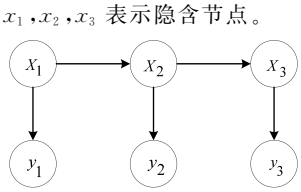


图 2 隐马尔可夫 HMM 的图形模式

Fig. 2 Hidden Markov model diagram

HMM 的数学表达式可简化为：$\lambda=(A,B,\pi)$。现将各符号含义说明如下：$Q=q_1,q_2,\cdots,q_T$ 表示隐状态序列；$O=o_1,o_2,\cdots,o_T$ 表示可输出观测序列；$\boldsymbol{A}=a_{11}a_{12}\cdots a_{m}$ 表示转移概率矩阵，a_{ij} 表示从状态 i 转移到状态 j 的概率且需满足式(7)的要求；$B=b_i(o_t)$ 表示发射概率矩阵，即在状态 i 观测到 o_t 的概率；$\pi=\pi_1,\pi_2,\cdots,\pi_n$ 表示初始状态，π_i 表示初始状态为 i 的概率且满足式(8)。

$$\sum_{j=1}^{n}a_{ij}=1 \tag{7}$$

$$\sum_{i=1}^{n}\pi_i=1 \tag{8}$$

构建隐马尔可夫航班延误预测分析模型，采用解码问题 Viterbi 算法来对模型进行构建和分析。Viterbi(维特比算法)就是求解 HMM 上的最短路径，也是最大概率问题的算法，根据可观察状态的序列找到一个最可能的隐藏状态序列。已知一个观测值序列 $O=o_1,o_2,\cdots,o_T$，同时也已知隐马尔可夫模型 $\lambda=(\pi,A,B)$，尝试通过数学方法选取一组与观测值序列 $O=o_1,o_2,\cdots,o_T$ 对应的隐状态序列 $Q=q_1,q_2,\cdots,q_T$。通常认为使得式(9)取得最大的概率值即为最为合适的隐状态序列[12]。

$$\begin{aligned}P(O|\lambda)&=\sum P(O|Q,\lambda)P(Q|\lambda)\\&=\sum_{q_1q_2\cdots q_T}\pi_{q_1}b_{q_1}(Q_1)a_{q_1q_2}(Q_2)\cdots a_{q_{T-1}q_T}b_{q_T}(Q_T)\\P(O|\lambda)&=\sum P(O|Q,\lambda)P(Q|\lambda)\\&=\sum_{q_1q_2\cdots q_T}\pi_{q_1}b_{q_1}(Q_1)a_{q_1q_2}(Q_2)\cdots a_{q_{T-1}q_T}b_{q_T}(Q_T)\\P(O|\lambda)&=\sum P(O|Q,\lambda)P(Q|\lambda)\\&=\sum_{q_1q_2\cdots q_T}\pi_{q_1}b_{q_1}(Q_1)a_{q_1q_2}(Q_2)\cdots a_{q_{T-1}q_T}b_{q_T}(Q_T)\end{aligned} \tag{9}$$

首先定义 $\gamma_t(i)=P(q_t=S_i|O,\lambda)$ 是观测序列 O 和给定模型参数 λ，t 时刻恰为状态 S_i 的概率：

$$\gamma_t(i)=P(q_t=S_i|O,\lambda)=\frac{P(q_t=S_i,O|\lambda)}{P(O|\lambda)}=\frac{P(O_1,O_2,\cdots,O_t|q_t=S_i,\lambda)P(O_{t+1},O_{t+2},\cdots,O_T|q_t=S_i,\lambda)}{P(O|\lambda)} \tag{10}$$

再根据前向变量 $\alpha_t(i)$ 和后向变量 $\beta_t(i)$ 的定义，可将上式变形得到以下等式：

$$\gamma_t(i)=P(q_t=S_i|O,\lambda)=\frac{P(q_t=S_i,O|\lambda)}{P(O|\lambda)}=\frac{P(O_1,O_2,\cdots,O_t|q_t=S_i,\lambda)P(O_{t+1},O_{t+2},\cdots,O_T|q_t=S_i,\lambda)}{P(O|\lambda)}=\frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)}=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{11}$$

其中，$\gamma_t(i)$ 满足条件 $\sum_{i=1}^{N}\gamma_t(i)=1$，使得 $\gamma_t(i)$ 最大的状态即是最佳状态。

3 针对指定航班计划的延误预测

3.1 基于贝叶斯网络的航班离港延误数据分析

研究选取国内某大型航空公司航班运行历史时刻表，构建 BN 模型样本选取了 2019 年 1 月 1 日至 1 月 31 日共计 10698 个航班数据。因在民航局发布的《民航航班正常统计办法》中对航班延误的定义主要针对离港延误时间来界定，故本文重点分析航班离港延误的情况。首先数据预处理在 SQL Server 2008 环境中增加离港延误时间属性。根据航空公司航班运行时刻表中对延误原因的分类标注，归纳延误原因的标记代码如表 1 所列。

表 1 航班延误原因与延误代码对照表

Table 1 Flight delay reasons and delay codes

延误原因	代码	延误原因	代码	延误原因	代码
天气原因	TQ	流量控制	LL	旅客	LK
工程机务	JW	军事活动	JS	公共安全	AQ
公司计划	JH	机场	JC	代理机构	DL
空勤组	KQ	联检	LJ	地面服务	DM
食品供应	SP	离港系统	XT	其他说明	SM

从航班信息时刻表中筛选出离港航班信息并输入模型变量中，模型输出形式为 BN 的后验概率分布，建立航班数据的 BN 分析模型如图 3 所示。

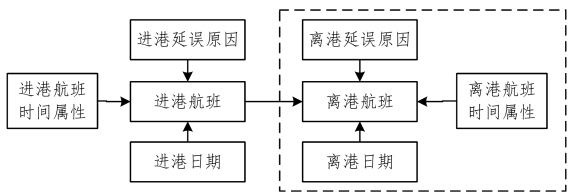


图 3 航班延误贝叶斯网络模型

Fig. 3 Bayesian network flight delay model diagram

利用 Netica 软件来分析的离港航班数据 BN 模型[13]，变量集为：$X=\{cur\_date,time,delay\_code,dep\_delay\}$。其中，$cur\_date$ 表示当前日期；$time$ 表示一天当中的时间段；$delay\_code$ 表示离港延误原因；dep_delay 表示离港延误的时长。

各个变量的值域为：$cur_date=\{$D20190101,D20190102,D20190103,$\cdots$,D20190131$\}$。航班历史运行数据 2019 年 1 月份每一天的航班离港量。

$time=\{0\sim8,8\sim10,10\sim12,12\sim16,16\sim18,18\sim20,20\sim24\}$ 将一天 24 小时划分为这 7 个时间段。

$delay_code=\{$天气原因、工程机务、公司计划、空勤组、食品供应、流量控制、军事活动、机场、联检、离港系统、旅客、公共安全、代理机构、地面服务、其他说明$\}$。(对应表 1)

$dep_delay=\{$小于 30 min,30～60 min,1～2 h,2～4 h,大于 4 h$\}$。时间划分依据民航总局颁发的《民航航班正常统计办法》的通知中航空公司航班延误时间统计表中的时段划分来确定的。

从所建立的航班延误 BN 模型中可以清楚地看出，航班延误时间 dep_delay 的父节点有两个，分别为 $time$ 和 $delay\_code$，由此表示出了其与计划离港时间和延误代码之间的关系。因此，通过从航班运行数据出发的贝叶斯网络来分析离港航班延误的延误原因与延误时间，挖掘其中的关联性并做分析。

不管是航班延误原因、航班计划离港时段、已知机场离港延误情况，每个变量的状态都对其他状态的发生有一定的影响。获取上述影响因素联合分布的条件概率过程，实质上就是航班延误贝叶斯网络模型学习的过程。当新的航班信息数据加入之后，之前得到的后验概率就成为了先验概率，进而又得到了新的后验概率分布，这样将大量数据加入航班延误数据 BN 模型之后，各节点边缘概率分布就趋于一个相对平稳的概率分布。

t0 to t8 表示计划离港航班时间在 0 点至 8 点之间的航班量占全部航班总量的 8.73%。各节点的条件概率分布以节点 dep_delay 的条件概率分布部分内容为例进行说明，如表 2 所列。

表 2 dep_delay 概率分布 CPT(部分内容)

Table 2 dep_delay probability distributions

time	delay_code	<30 min	30～60 min	1～2 h	2～4 h	>4h
0－8 点	TQ	11	6	18	45	54
0－8 点	JW	0	1	3	1	0
0－8 点	JH	1	8	4	9	2
0－8 点	KQ	0	0	0	1	1
0－8 点	SP	0	0	0	0	0
0－8 点	DL	0	0	0	0	0
0－8 点	DM	0	0	1	0	0
0－8 点	LL	16	25	21	12	3
0－8 点	SM	471	3	1	2	8

通过图 4 可以看到离港航班延误的整体概率分布情况，清楚地描述了贝叶斯网络模型中 4 个节点随机变量的贝叶斯后验概率分布，进一步分析出了延误时间在 30 min 以内的航班占了绝大部分的 57.3%，延误时间在 30～60 min 之间的占到 17.8%，1 小时至 2 h 的航班延误有 11.4%，同时 2～4 h 延误以及超过 4 h 的延误也有发生，但发生的概率很小。另外，可以从模型中发现导致航班延误的主要原因为流量控制(LL)，贝叶斯后验概率占到 20.9%，天气原因也是导致延误的主要因素，后验概率为 18.9%，说明这段时间天气状况不是很稳定，此外有 6.58%是公司计划的原因。以小概率存在由于机场和旅客自身原因导致的航班延误因素后验概率分别为 1.18%和 1.03%。

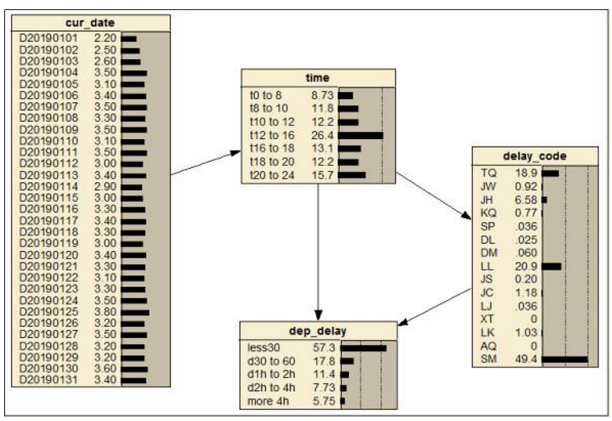


图 4 离港航班延误 BN 模型

Fig. 4 Bayesian network of flight delays diagram

当大量新的航班 BN 模型数据不断加入来更新后验概率分布并趋于稳定的情况下可去除日期变量，并将这一类环境下的延误情况作为目标时段内延误的参考。通过图 4 可以很直观地分析到在时段 12:00－16:00 之间延误情况最为严重，图 5 给出在此时间段内航班离港延误的情况，可以看到绝大部分航班延误时间在 30 min 以内，另外也存在一定比例的 30～60 min 的延误和 1～2 h 的延误，导致延误发生的主要原因除了前面已做过说明分析的 SM(其他说明原因)以外，依次为 LL(流量控制原因)和 TQ(天气原因)，比例几乎持平后验概率分别为 20.9%和 20%，其次就是因为 JH(航空公司计划)的原因导致的延误发生，后验概率为 8.5%。

进一步分析 30～60 min 航班延误的原因，从图 6 中可以看出结论与图 5 基本一致，主要延误原因是 LL(流量控制)、JH(公司计划)和 TQ(天气原因)，但概率分布不同，30～60 min时间段上的延误主要以 49.9%的概率由 LL(流量控制)原因导致延误的发生。天气原因并不是导致此时间段延误的主要原因，可以推测此时段天气状况较为稳定。

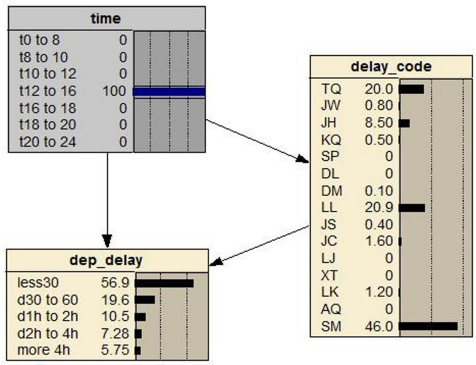


图 5 12:00－16:00 离港延误情况

Fig. 5 12:00－16:00 departure delay diagram

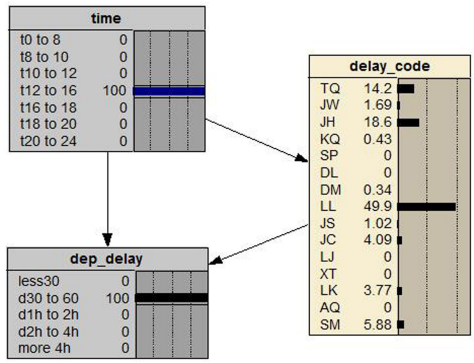


图 6 30～60 min 离港延误原因

Fig. 6 30～60 min departure delay diagram

通过图 7 可以明显的分析出，对于延误时长在 1～2 h 之间的延误仍然符合延误的主要原因是 LL(流量控制原因)和 TQ(天气原因)的结论，但是 TQ(天气原因)是导致该时段发生延误的最主要原因，后验概率分布高达 46.7%。其次是 LL(流量控制)原因占 34.4%的概率，说明导致 1～2 h 的延误天气原因起到了主导作用。

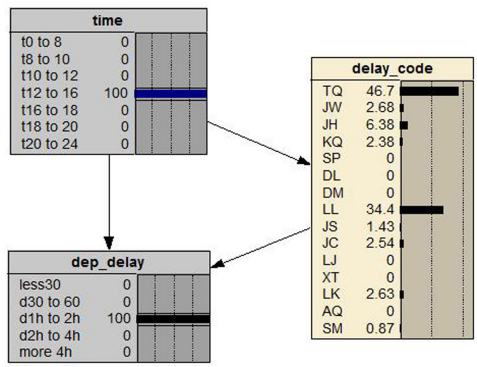


图 7 1～2 h 离港延误原因

Fig. 7 1～2 h departure delay diagram

天气原因是导致样本数据 1 月份航班离港延误的一个主要因素，对天气因素导致的延误影响情况是很值得关注的。图 8 给出了 TQ(天气原因)延误影响情况的概率分布，发现在延误高发的时段 12:00－16:00 之间，延误时长多集中在 2～4 h，后验概率为 29.3%。同时延误时长大于 4 h 的概率占到 26.6%，1～2 h 的概率占到 24.6%，可见天气导致的延误时间较长，该时段因天气原因导致的延误情况较为严重，并且波及范围较广，需要给予特别关注。

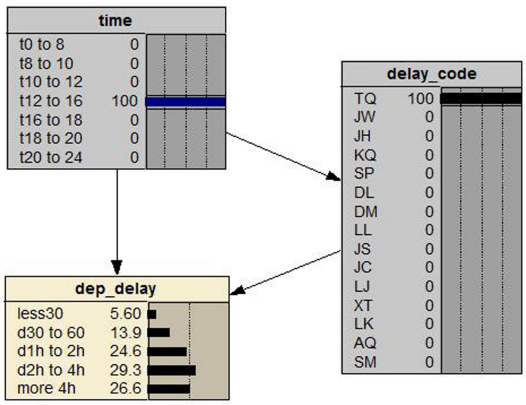


图 8 TQ(天气)原因导致延误的情况

Fig. 8 TQ (weather) caused delays diagram

下面通过分析 LL(流量控制)原因在延误高发时间段 12:00－16:00 之间的延误情况,如图 9 所示。可以从后验概率分布情况看出,LL 原因对航班延误时间影响主要集中在 30～60 min 之间,后验概率占到 46.7%,其次造成延误小于 30 min 的后验概率为 32.5%,造成 2 h 以上的航班延误情况很少。

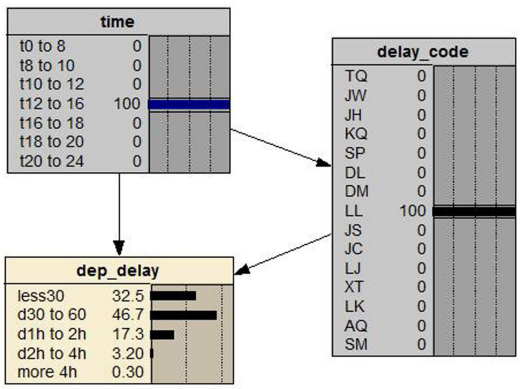


图 9 LL(流量控制)原因导致延误情况

Fig. 9 LL (flow control) caused delays diagram

在分析离港航班延误数据 BN 模型时还有一部分 JH(公司计划)原因导致的航班延误,图 10 特别针对 JH(公司计划原因)导致的延误情况进行展示,在延误高发的 12:00－16:00 时间段,JH 原因导致的延误影响集中发生在小于 30 min 的延误时长,后验概率为 47.1%;30～60 min 的延误时长,后验概率为 42.9%;1～2h 的延误后验概率占到7.9%;另外极少概率为导致 2h 以上的离港延误。

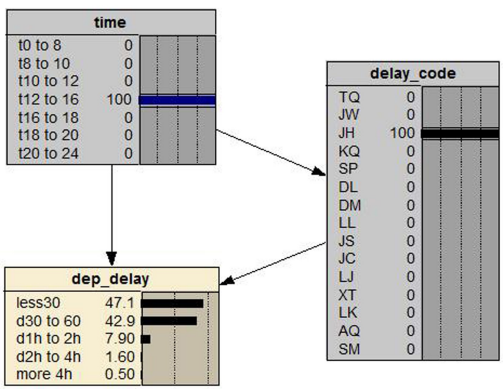


图 10 JH(公司计划)原因导致延误的情况

Fig. 10 JH (airline schedule) caused delays diagram

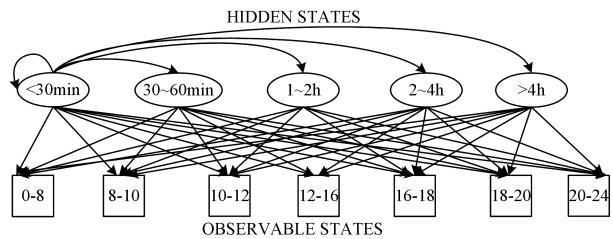
综上所述,可以通过不同条件下的航班延误概率分布情况了解造成航班延误的主要原因和影响航班延误的时间范围。通过概率分布情况基本可以得到数据之间的隐藏关系为天气原因造成航班超过 2 h 以上的延误可能性最大,另外流量控制原因和公司计划原因最有可能造成航班延误时间在 1 h 以内。

3.2 航班延误隐马尔可夫预测模型

通过前文构建的航班延误贝叶斯网络模型的 BN 结构进行的相关数据分析可知,航班延误时间和航班延误的原因是影响航空公司整体运行品质的重要因素,也是航空公司决策者比较关心的两个因素,因此这里针对构建航班延误隐马尔可夫模型构建过程从延误时间以及延误原因两个方面进行建模分析。

(1)计划离港时间作为输出可观测序列,预测可能发生的延误时间。

针对图 11,需要特别注意的是该模型预测的对象为某架飞机所执行的一系列航班,且该系列航班的起飞机场为已知,计划离港时间是存在时序关系的。所建立的 HMM 模型利用 Viterbi 算法便可以对其进行相应的延误时间预测。



注:该图只针对<30 min 的节点做了转移概率示意弧线,其他节点应同理

图 11 航班计划离港时间为输出观测序列的 HMM 预测模型

Fig. 11 HMM prediction model with flight planned departure time as output observation sequence diagram

HMM 模型使用的条件为观测序列和隐状态序列都必须是离散的,因此将隐状态序列离散化处理如表 3、表 4 所列。

表 3 隐状态变量离散化

Table 3 Discretization of hidden state variables

隐状态变量:延误时间	离散化序列表示
<30 min	1
30～60 min	2
1～2h	3
2～4h	4
>4h	5

表 4 可观测状态变量(计划离港时间)离散化

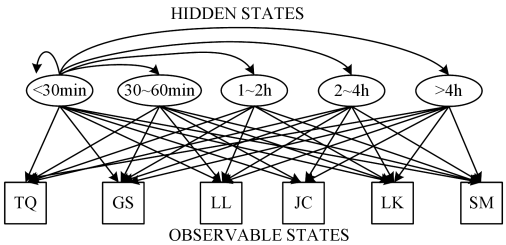
Table 4 Discretization of observable state variable (planned departure time)

可观测状态变量:计划离港时间	离散化序列表示
0:00－8:00	1
8:00－10:00	2
10:00－12:00	3
12:00－16:00	4
16:00－18:00	5
18:00－20:00	6
20:00－24:00	7

(2)航班延误原因作为可观测输出序列，预测可能发生的延误时间。

当航空公司得知航班确定已延误时通常已经确定了该航班的延误原因，但是该航班延误的时间大概有多长是困扰航空公司非正常航班预警的一个关键问题。

图 12 给出了建立的航班延误原因作为可观测序列的 HMM 预测模型，其中隐状态序列依然为航班的延误时间。所建立的 HMM 模型利用 Viterbi 算法便可以对其进行相应的延误时间预测。



注：为图示清晰，只针对<30 min 的节点做了转移概率示意弧线，其他节点应同理

图 12 延误原因为输出可观测序列的 HMM 预测模型

Fig. 12 HMM prediction model with flight delay reasons as output observation sequence diagram

同样需要对可观测序列航班延误原因状态变量进行离散化处理，这里为了便于模型的建立，将延误原因按照航空公司延误原因统计最多的六大类来处理，如表 5 所列。

表 5 可观测状态变量：延误原因离散化

Table 5 Observable state variables: discretization of the causes of delay

可观测状态变量：延误原因	离散化序列表示
TQ	1
GS	2
LL	3
JC	4
LK	5
SM	6

4 数值分析

针对所建立的航班延误 HMM 预测模型，根据计划离港时间为可观测输出序列给出仿真实验。设置计划离港时间为可观测输出序列的 HMM 预测模型，模型训练数据来源选取某大型航空公司 2019 年 1－2019 年 6 月份的航班历史运行数据。

首先航班时刻表数据信息需要在 SQL Server 2008 环境中进行预处理，筛选出某指定机型所对应的一系列航班。如表 6 中筛选出的机号为 6779，并且起飞站为成都所对应的所有航班。本算例以预测 7 月份连续两天机号 6779 所执行起飞站为成都的航班为例，经数据库筛选得到计划离港时间，并结合表 4 离散化处理对照序列表可得到如表 7 所列的可观测输出状态序列表。

表 6 某航空公司 2019 年机号 6779 运行信息汇总(部分截取)

Table 6 Operating information tail number 6779 in 2019

航班号	机型	机号	起飞站	计划起飞	实际起飞	到达站	延误时间	延误原因
AB8555	A320	B6779	成都	7:15	7:20	西安	—	NULL
AB8643	A320	B6779	成都	12:50	13:11	西昌	0:16	军事活动
AB8895	A320	B6779	成都	17:15	17:57	北京	0:38	军事活动
AB8733	A320	B6779	成都	20:30	21:13	广州	0:38	军事活动
AB8643	A320	B6779	成都	7:10	7:22	西昌	0:07	军事活动
AB8895	A320	B6779	成都	11:45	12:44	北京	0:59	航空公司
AB8896	A320	B6779	成都	16:05	16:41	西昌	0:31	机场
AB8733	A320	B6779	成都	20:35	21:11	广州	0:31	机场
AB8751	A320	B6779	成都	21:00	23:57	海口	2:52	天气
AB8985	A320	B6779	成都	8:00	13:49	武汉	5:44	天气
AB8733	A320	B6779	成都	12:50	15:48	广州	2:53	天气
AB8871	A320	B6779	成都	7:15	8:43	太原	1:23	天气
AB8919	A320	B6779	成都	20:30	22:15	杭州	1:40	天气
AB8667	A320	B6779	成都	21:50	21:48	昆明	—	NULL
AB8723	A320	B6779	成都	7:35	8:13	呼和浩特	0:33	流量

表 7 计划离港时间离散状态序列表

Table 7 Discrete sequence of planned departure time

计划离港时间	07:15	12:50	17:15	20:30	07:10	11:45	16:05	20:35
离散化序列	1	4	5	7	1	3	5	7

根据已完成的航班离港延误 BN 模型数据分析，可以得到延误时间的隐状态转移概率矩阵 $\boldsymbol{a}$ 如表 8 所列。

表 8 航班延误时间转移概率矩阵分布

Table 8 Probability matrix distribution of flight delay

$\boldsymbol{a}$	<30 min	30～60 min	1～2 h	2～4 h	>4 h
<30 min	0.6	0.1	0.1	0.1	0.1
30～60 min	0.1	0.5	0.2	0.1	0.1
1～2 h	0.1	0.1	0.4	0.2	0.2
2～4 h	0.1	0.1	0.1	0.4	0.3
>4 h	0.1	0.2	0.2	0.2	0.3

表 8 中的纵列表示当前状态，横行表示未来预测状态，以表格中第一个数字 0.6 为例，即当前已知延误时间<30min，未来预测延误时间仍然<30min 的概率为 0.6。根据航班时刻表 1－6 月份的历史数据 90%作为模型训练数据统计，得到计划离港时间的可见状态发生输出概率分布 b_1 如表 9 所列。

表 9 计划离港时间输出概率矩阵分布

Table 9 Output probability matrix distribution of planned departure time

b_1	0:00－8:00	8:00－10:00	10:00－12:00	12:00－16:00	16:00－18:00	18:00－20:00	20:00－24:00
<30 min	0.103	0.144	0.144	0.261	0.114	0.109	0.125
30～60 min	0.029	0.131	0.112	0.292	0.147	0.132	0.156
1～2 h	0.054	0.083	0.071	0.245	0.166	0.140	0.240
2～4 h	0.110	0.028	0.123	0.250	0.132	0.141	0.217
>4 h	0.140	0.000	0.031	0.264	0.180	0.153	0.231

初始状态设置为1，可见状态序列 $V=\{1,4,5,7,1,3,5,7\}$（即表7所列计划离港时间离散状态序列表）的MATLAB运行结果如图13所示。

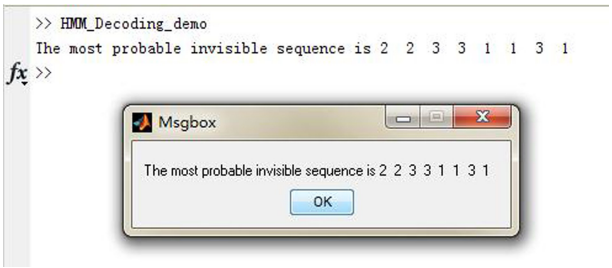


图13 航班延误预测Viterbi算法的MATHLAB运行结果

Fig. 13 Flight delay prediction Viterbi algorithm MATHLAB operation result

结合表3所列的航班延误时间隐状态离散处理结果，对应得出预测的航班延误时间的离散序列为{2 2 3 3 1 1 3 1}，分别表示预测延误时间依次为{30～60 min，30～60 min，1～2 h，1～2 h，<30 min，<30 min，1～2 h，<30 min}。

现基于该方法，选取机号6779，离港站为成都机场，利用MATLAB实现预测7月份的所有航班延误结果如图14所示，对比信息整理结果如表10所列。

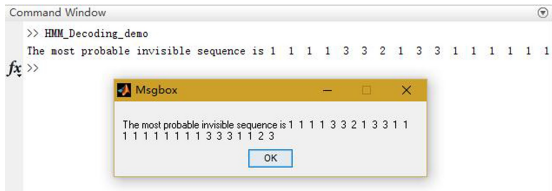


图14 7月份延误预测Viterbi算法的MATLAB运行结果

Fig. 14 Delay forecast Viterbi algorithm MATLAB running results in July

表10 某航空公司7月份离港延误预测情况与实际情况的对比统计

Table 10 Statistics comparing the forecast of departure delay with actual situation in July

序号	航班日期	航班号	机号	起飞站	计划离港时间	到达站	计划离港时间离散化	预测结果离散化数值	预测结果	实际延误时间
1	2019/7/2	8555	6779	成都	7:00	西安	1	1	<30 min	7
2	2019/7/2	8801	6779	成都	16:00	西安	5	1	<30 min	137
3	2019/7/4	8555	6779	成都	7:00	西安	1	1	<30 min	0
4	2019/7/4	8643	6779	成都	16:05	西昌	5	1	<30 min	27
5	2019/7/4	8895	6779	成都	20:35	北京	7	3	1～2 h	10
6	2019/7/5	8801	6779	成都	16:00	西安	5	3	1～2 h	5
7	2019/7/7	8733	6779	成都	12:50	广州	4	2	30～60 min	27
8	2019/7/7	8871	6779	成都	7:15	太原	1	1	<30 min	8
9	2019/7/7	8895	6779	成都	20:35	北京	7	3	1～2 h	5
10	2019/7/8	8627	6779	成都	15:40	无锡	5	3	1～2 h	9
11	2019/7/8	8896	6779	成都	11:45	西昌	3	1	<30 min	41
12	2019/7/9	8101	6779	成都	7:00	南充	1	1	<30 min	29
13	2019/7/13	8755	6779	成都	17:15	三亚	5	1	<30 min	14
14	2019/7/13	8997	6779	成都	7:25	长沙	1	1	<30 min	21
15	2019/7/18	8631	6779	成都	15:10	昆明	4	1	<30 min	23
16	2019/7/18	8923	6779	成都	8:55	南京	2	1	<30 min	8
17	2019/7/21	8813	6779	成都	15:15	徐州	4	1	<30 min	30
18	2019/7/22	8631	6779	成都	15:10	昆明	4	1	<30 min	146
19	2019/7/22	8923	6779	成都	8:55	南京	2	1	<30 min	23
20	2019/7/24	8617	6779	成都	15:00	香港	4	1	<30 min	28
21	2019/7/24	8667	6779	成都	21:50	昆明	7	1	<30 min	21
22	2019/7/28	8867	6779	成都	19:20	西宁	6	3	1～2 h	1
23	2019/7/28	8868	6779	成都	23:55	昆明	7	3	1～2 h	5
24	2019/7/29	8855	6779	成都	10:05	银川	3	1	<30 min	26
25	2019/7/29	8856	6779	成都	15:05	昆明	4	1	<30 min	23
26	2019/7/29	8867	6779	成都	19:20	西宁	6	2	30～60 min	35
27	2019/7/29	8868	6779	成都	23:55	昆明	7	3	1～2 h	95

通过对7月份一整个月的机号6779并且起飞站为成都的离港航班延误预测情况进行统计分析，其中延误预测准确的有18个航班，误差率为33.3%，可见预测模型是可行且有效的，预测效果较为理想。

结束语 本文针对指定航班计划的延误预测分析，预测对象为某指定机号且离港机场为同一机场的航班计划。应用贝叶斯网络理论建立了实际航班数据的BN模型，利用Netica软件进行数据处理，挖掘各要素之间的关联性，给出了不同条件下航班延误的概率分布情况；建立了实际航班数据的隐马尔可夫航班延误预测分析模型，利用解码问题Viterbi算法实现了对指定航班延误时间的预测，提出了以计划离港时间为输出观测序列、以延误原因为输出观测序列两种不同条件下的隐马尔可夫（HMM）航班延误预测模型，通过实例仿真过程得出了延误时间的预测结果，可为以后研究预防以及控制航班长时间延误的措施提供理论依据。

参 考 文 献

[1] FABDELGHANY K, SHAH S S, RAINA S, et al. A Model for Projecting Flight Delays during Irregular Operation Conditions [J]. Journal of Air Transport Management (S0098-1856), 2004, 10(6): 385-394.

[2] TU Y F, BALL M O, JANK W S. Estimating Flight Departure Delay Distributions — A Statistical Approach with Long-Term Trend and Short-Term Pattern[Z]. Smith School Research Paper No. RHS 06-034, 2005.

（下转第485页）