

文章编号: 1001-8360(2004)05-0001-07

铁路客运量数据挖掘预测方法及应用研究

王艳辉, 王 卓, 贾利民, 秦 勇

(中国铁道科学研究院 电子计算技术研究所, 北京 100081)

摘 要: 在分析铁路客票数据特征的基础上, 提出采用分段模糊 BP 神经网络对铁路客运量进行数据挖掘预测。通过对铁路客票数据的分段处理, 提高了网络学习的收敛速度和预测精度, 并在 MATLAB 环境下建立了分段模糊 BP 神经网络模型, 在仿真试验中各分段的期望输出和实际输出之间吻合较好, 从而证明了分段模糊的数据处理方法是有效的; 同时, 预测的客运量和实际客运量数值非常接近, 说明分段模糊 BP 神经网络得到的数据挖掘预测模型对铁路客运量有很好的预测效果, 该预测模型可信, 为预测铁路客运量提出了一种新思路。

关键词: 铁路客票数据; 数据挖掘; 分段模糊 BP 神经网络; 旅客发送量

中图分类号: U492.413 **文献标识码:** A

Study on Prediction Method of Data Mining of the Passenger Traffic Volume of Railways and Its Application

WANG Yan-hui, WANG Zhuo, JIA Li-min, QIN Yong

(Research Institute of Computer Technology, China Academy of Railway Sciences, Beijing 100081, China)

Abstract: The segment fuzzy BP Neural Network is adopted to predict the passenger traffic volume of railways in data mining based on analyzing the data feature of railway passenger tickets. The convergence speed and prediction precision of the neural network are both improved by segmentation of the passenger ticket data. The model of the segment fuzzy BP Neural Network is set up using Matlab software. The emulation results show that the anticipant and the actual output results of all the segments are very similar, which proves that data processing of segment fuzzy is effective; at the same time, the anticipant and the actual output results of the passenger traffic volume are very proximate, which shows that the prediction model of data mining is reliable to predict the passenger traffic volume. This research provides a new thought to forecast the passenger traffic volume in data mining.

Key words: passenger ticket data of railways; data mining; segment fuzzy BP Neural Network; passenger traffic volume

近年来,随着铁路客票系统的大面积推广应用,目前已经积累了大量的客票发售数据,同时铁道部也建立了数据仓库^[1]。通过数据展现使用户动态掌握售票情况、实现指标统计和报表自由生成、进行灵活的多维分析^[2],而进一步的工作则是进行与空间、时间相关的统计技术分析,对铁路现有信息系统提供的数据进行综合分析和数据挖掘。根据分析、挖掘结果进行预测,为铁路旅客运输组织、运力调配提供决策支持,提高经济效益。

铁路客票数据具有明显的空间属性特征,空间数据的复杂性和非线性对一般的数据挖掘方法提出了新的挑战,研究面向具有空间属性系统的数据挖掘方法具有重要意义。为此,本文以铁路假日运输管理系统中春运期间的客票数据为依据,提出一种基于分段模糊 BP 神经网络方法,对春节前 19 天至节后 25 天的客运量进行预测。

1 铁路客票的数据特征

铁路客票数据的复杂性特征主要表现在以下 5 方面:

(1) 数据量大

收稿日期: 2004-05-09; 修回日期: 2004-06-03

作者简介: 王艳辉(1974—), 男, 河北保定人, 助理研究员, 博士后。

E-mail: zhxywyh@sohu.com

大量的数据常使一些方法因算法难度或计算量过大而无法得以实施,因而数据挖掘的任务之一就是要创建新的计算策略,并发展新的高效算法克服由数据量大造成的技术困难。

(2) 客票数据空间属性之间的非线性关系

客票空间属性之间的非线性关系是客票数据具有空间系统复杂性的重要标志,其中蕴含着系统内部作用的复杂机制,因而在数据挖掘时被作为客票空间数据挖掘的主要任务之一。

(3) 客票空间数据的尺度特征

客票空间数据的尺度特征是指空间数据在不同观察层次上所遵循的规律以及体现出的特征不尽相同。尺度特征是空间数据复杂性的又一表现形式,利用该特征可以探究空间信息在概化和细化过程中所反映出的特征渐变规律。

(4) 客票空间信息的模糊性

客票数据复杂性的另一个特征就是空间信息的模糊性。模糊性几乎存在于各种类型的空间信息中,如空间位置的模糊性、空间相关性的模糊性以及模糊的属性值等。

(5) 客票数据的缺值

缺值现象源自某种不可抗拒的外力而使数据无法获得或发生丢失。如何对丢失数据进行恢复并估计数

据的固有分布参数,成为解决客票数据复杂性的难点之一。

铁路客票数据所表现出的上述复杂性特征为相应的数据挖掘研究提出了更高要求。

2 铁路客票数据挖掘的建模方法

由于客票数据属性的特殊性和复杂性,面向具有空间属性的客票数据挖掘需要解决以下 3 方面的问题:

(1) 客票数据模型的建立 怎样通过建模的方法把客票的空间属性和其数据属性有机的融合在一起。

(2) 数据挖掘 怎样解决客票数据的高维问题;怎样在数据挖掘时考虑空间位置关系。

(3) 知识的表达 如何直观的表达空间个体的空间关系。

鉴于此,将采用分段模糊 BP 神经网络的数据挖掘方法进行铁路客票数据预测研究。通过分段模糊和神经网络的建模方法将客票的空间属性和其数据属性有机地结合起来,在充分考虑空间关系的基础上解决数据的高维问题,将系统的结构含于网络的数值中,使其具有强大的学习能力和适应能力,有效的解决空间数据的复杂性和非线性问题。

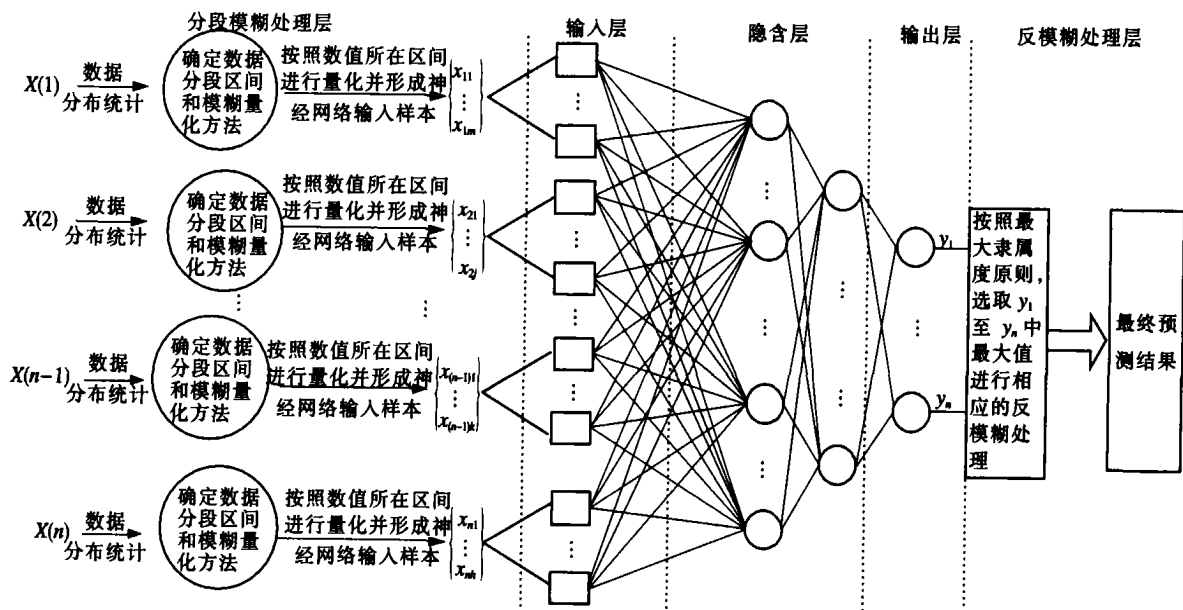


图 1 分段模糊 BP 神经网络结构

2.1 分段模糊 BP 神经网络

将模糊系统与 BP 神经网络进行有机融合,可以有效地发挥其各自的优势并弥补其不足。模糊数学和

神经网络的融合形态多种多样,其特点各异,适合于不同的应用情况^[3]。现提出一种串形结构的分段模糊 BP 神经网络。该网络由 3 部分组成,前一部分为分段

模糊量化部分,第二部分为神经网络学习部分,第三部分分为反模糊数据处理阶段,如图 1 所示。

分段模糊量化部分主要是根据输入因素和输出因素的数据分布区间或定性分类进行分段模糊处理,处理后的数据样本遵循模糊数学中最大隶属度的原则进行相应的排列(见实例)。神经网络输出层输出的数据则按照最大隶属度原则进行相应的反模糊处理得出最终的训练或预测结果。

2.2 客票数据分段模糊 BP 神经网络的建模

2.2.1 学习样本的选取及预测目标

以铁路假日运输管理系统中春运期间的客票数据为依据,以北京、郑州、济南、呼和浩特和沈阳 5 个铁路局 1994~2002 年春节前 19 天到节后 25 天每天直通旅客发送量(以下称发送量)作为学习样本,共计 405 个学习样本;以上述 5 个铁路局的 2003 年相应的数据作为预测数据,预测目标为:

- (1) 客流预测的时间范围是春运期间,从节前 19 天到节后 25 天;
- (2) 客流预测的时间间隔为天;
- (3) 预测对象是旅客发送量;
- (4) 预测是宏观预测,对铁路局的旅客发送量进行预测;
- (5) 数据来源是 18 点系统。

2.2.2 分段模糊 BP 神经网络的建模

对于铁路局间的旅客出行来说,距离较远的路局间的旅客出行没有距离较近的路局间的旅客出行频繁,邻近的路局间具有较大的关联性^[4]。为此,把北京路局邻近的沈阳、济南、呼和浩特和郑州路局组成空间序列,如图 2。

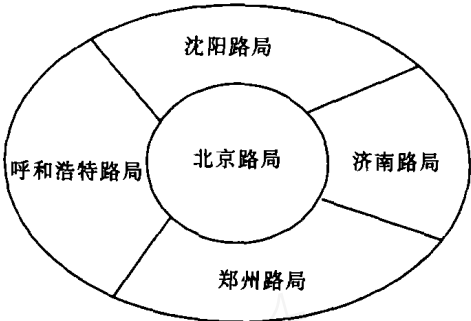


图 2 路局空间序列数据建模模型

在图 2 中,神经网络的输入因素为 4 个路局(郑州、济南、呼和浩特和沈阳)对应的同一天客流数据,而输出则对应北京路局相同时间的客流数据。

通过对 5 个路局 1994~2003 年的数据统计分析发现,各路局的旅客发送量相差悬殊,具体见表 1。

表 1 1994~2003 年 5 个铁路局客运数据统计 单位:人/天

统计项	铁路局名称				
	呼和浩特	郑州	济南	沈阳	北京
最大发送量	47 097	233 505	113 784	566 582	377 484
最小发送量	3 306	67 713	18 315	159 489	88 147

由表 1 可见,在 5 个路局中,呼和浩特路局的发送量最小,统计的最大发送量也仅为 47 097 人/天,而郑州、沈阳和北京路局的最小发送量都要比其最大发送量数量还要大,因此,由于路局间的数据过分悬殊,若直接采用神经网络进行学习,势必造成学习结果不收敛,必须采取一定的方法对数据进行处理。本文采用分段模糊方法对数据进行归一化处理,各路局数据分段,见表 2。

表 2 各铁路局数据的分段 单位:万人

路局名称	1	2	3	4	5	6	7	8	9	10
呼和浩特	0.1~0.5	0.5~1	1~1.5	1.5~2	2~2.5	2.5~3	3~3.5	3.5~4	4~4.5	4.5~5
郑州	5~7	7~9	9~11	11~13	13~15	15~17	17~19	19~21	21~23	23~25
沈阳	15~20	20~25	25~30	30~35	35~40	40~45	45~50	50~55	55~60	60~65
济南	1.5~2.5	2.5~3.5	3.5~4.5	4.5~5.5	5.5~6.5	6.5~7.5	7.5~8.5	8.5~9.5	9.5~10.5	10.5~11.5
北京	5~10	10~15	15~20	20~25	25~30	30~35	35~40			

按照上述的分段,采用升半梯形分布的函数对数据进行归一化处理^[5],其公式如下

$$\mu(x) = \begin{cases} 0 & x \leq a_1 \\ \frac{x - a_1}{a_2 - a_1} & a_1 < x \leq a_2 \\ 1 & x > a_2 \end{cases} \quad (1)$$

在此以 2000 年春节前第 13 天春运数据为例介绍学习样本的形成过程。该天呼和浩特、郑州、济南、沈阳和北京路局的发送量分别为 31 833、185 669、73 375、341 055、267 698。与表 2 中的分段区间相对照,如呼和浩特路局的 31 833 落在第 7 段内,则此时式(1)中的 $a_1 = 30\ 000$ 、 $a_2 = 35\ 000$,则式(1)变为

$$\mu(x) = \begin{cases} 0 & x \leq 30\ 000 \\ \frac{x - 30\ 000}{35\ 000 - 30\ 000} & 30\ 000 < x \leq 35\ 000 \\ 1 & x > 35\ 000 \end{cases} \quad (2)$$

将 $x = 31\ 833$ 代入式(2)得

$$\mu(31\ 833) = 0.366\ 6$$

其它也是如此,不再赘述。

其它区间内的样本输入值均取 0.000 1,这一方面提高了预测精度,减少了其它区间段数据的干扰,另一方面也加快了网络学习的收敛速度。计算后,该样本的输入为

0.000 1 0.000 1 0.000 1 0.000 1 0.000 1
0.000 1 0.366 6 0.000 1 0.000 1 0.000 1
0.000 1 0.000 1 0.000 1 0.000 1 0.000 1
0.837 5 0.000 1 0.000 1 0.000 1 0.000 1
0.000 1 0.000 1 0.000 1 0.821 1 0.000 1
0.000 1 0.000 1 0.000 1 0.000 1 0.000 1
0.000 1 0.000 1 0.000 1 0.000 1 0.000 1
0.000 1 0.783 4 0.000 1 0.000 1 0.000 1;

所对应的输出为

0.000 1 0.000 1 0.000 1 0.000 1 0.354 0
0.000 1 0.000 1。

3 网络参数设计及训练结果分析

3.1 网络参数设计

(1) 输入层节点与输出层节点的确定

输入层节点与输出层节点的个数往往是由具体问题而决定^[6]。在本文中,输入的因素为呼和浩特、济南、沈阳、郑州 4 个路局的同一天直通旅客发送量数

据,共 4 个因素,每个因素各有 10 个区段,共有 40 个区段,从而确定输入层的节点数为 40 个;又由于这些分级指标决定输出结果——北京路局的同一天直通旅客发送量,分为 7 个级别,所以输出层节点数为 7 个。

(2) 隐含层层数及节点数的确定

误差精度的提高有两种途径:一是增加隐含层的层数;二是增加隐含层节点个数。综合以上因素,并基于文献[7,8],本文采用具有 3 个隐含层的 BP 神经网络。对于如何确定最佳隐单元数的个数,目前尚无确定的办法,实际应用中要通过试算靠经验选取。

最佳节点数为

$$L_i = \sqrt{m + n + c_i} \quad (3)$$

式中, L_i 为第 i 隐含层的节点数; m 为输入节点数; n 为输出节点数; c_i 为介于 1~10 的常数。

这里 $m = 40$, $n = 7$,经过多次对比训练,选取 $c_1 = 3$,则 $L_1 = 10$; $c_2 = 9$,则 $L_2 = 16$; $c_3 = 7$,则 $L_3 = 14$ 。

(3) 初始权值的确定

取初始权值为 $(-1, 1)$ 之间的随机数。

(4) 学习速率的确定

一般选取适中的学习速率,选取范围在 0.01~0.8 之间^[9],取学习速率为 0.2。

(5) 学习动量的确定

动量法降低了网络对于误差曲面局部细节的敏感性,有效地抑制网络陷于局部极小,学习动量为 0.4。

(6) 期望误差的确定

期望误差值是通过不同期望误差网络的对比训练来选取的。经过多次对比训练,选取样本期望误差为 0.008。

3.2 网络训练结果及分析

网络训练在 Matlab 6.1 环境下进行。经过 56 062 步的学习运算后,系统达到了期望的学习误差。图 3 是网络学习到 1 000 步和 56 062 步时的误差曲线,图 4 是各样本输出节点的实际输出和期望输出的对比图。

从图 3 可以看出,网络在学习过程中没有产生局部振荡和陷入极小现象,整个学习过程中系统稳定性较好。从图 4(a)、(b)、(g)可以看出,期望输出和实际输出值主要分布在 0 附近,主要是由于发送量数据落在这 3 个区段的较少,大部分数据分布在其它 4 个区段内。总的来看,各分段的期望输出和实际输出之间吻合较好,从而证明该数据处理方法的有效性和网络学习参数的合理性。

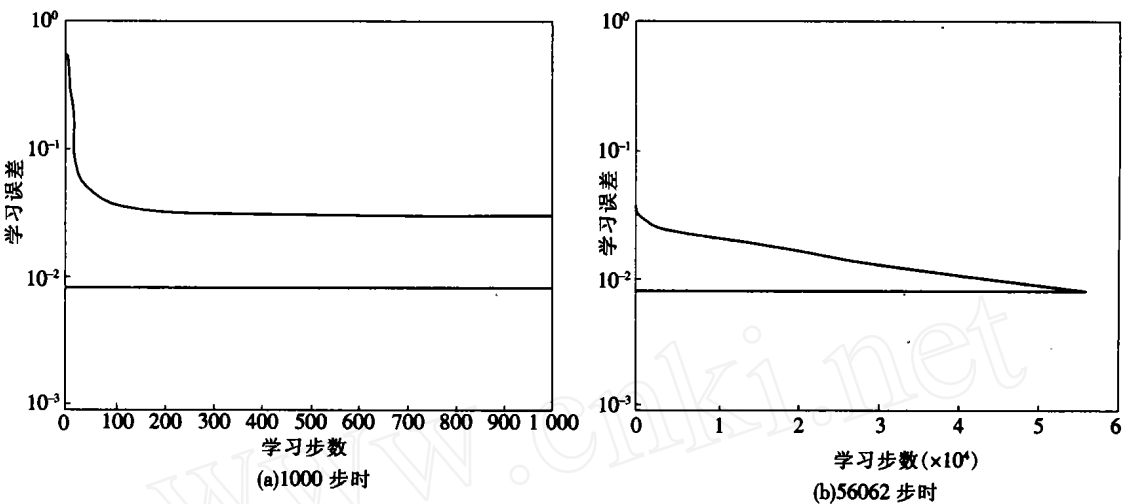
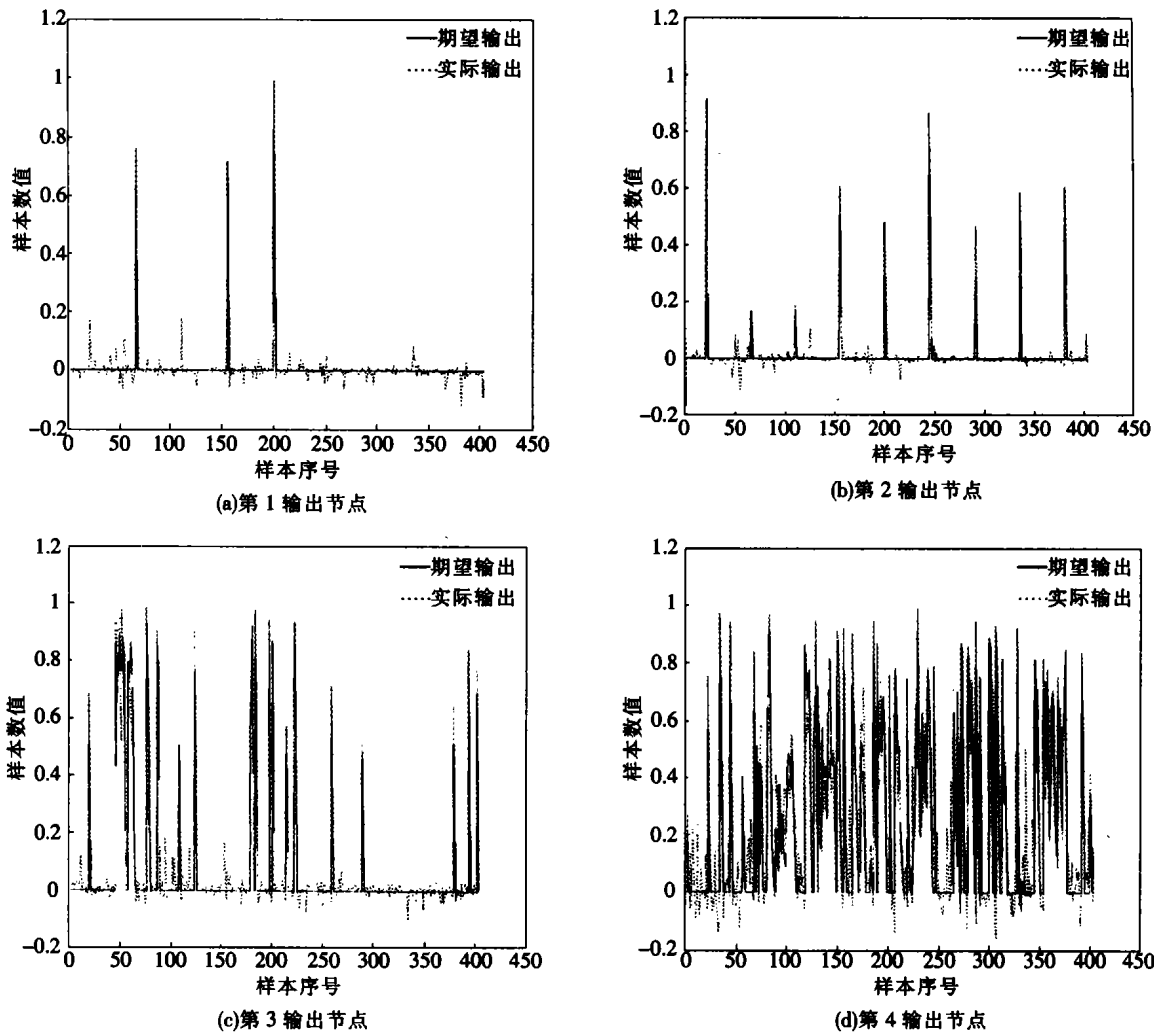


图 3 网络学习误差变化曲线



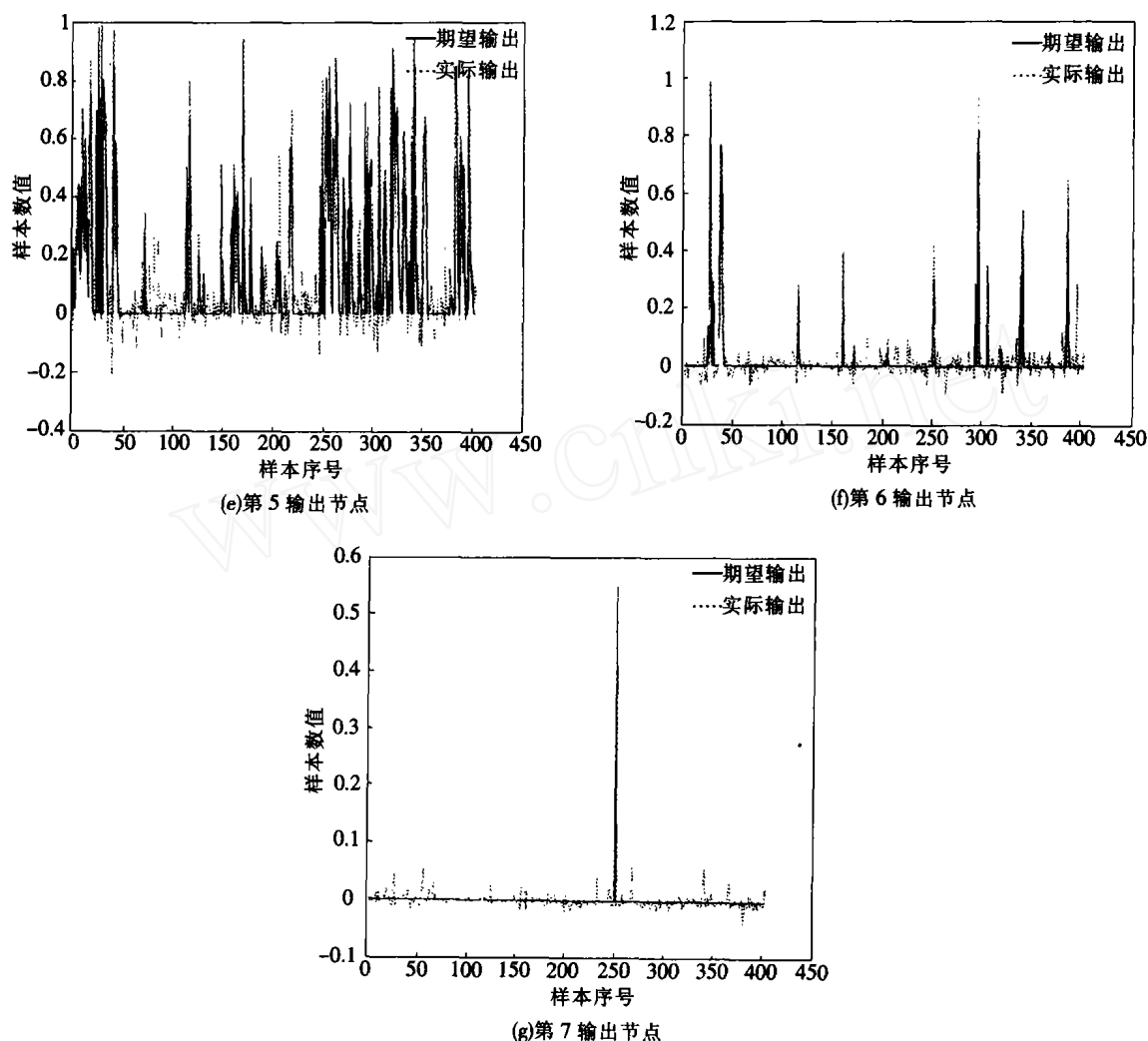


图 4 各样本输出节点的实际输出和期望输出

数值是比较接近的,从而进一步证明了网络学习的有效性和合理性。

4 分段模糊 BP 神经网络在铁路客票仿真预测中的应用

为了验证分段模糊 BP 神经网络所建模型的合理性和有效性,以 2003 年 4 个路局的实际直通旅客发送量为例,对 2003 年北京路局的直通旅客发送量进行预测,并与实际的直通旅客发送量进行比较。

数据的分段采取上述方式进行处理。对于 7 个输出值,则遵循模糊数学中最大隶属度原则,挑选样本输出中最大值进行反算得出真实的发送量。表 3 为北京路局实际直通旅客发送量与预测旅客发送量的比较。图 5 为 2003 年北京路局实际发送量与预测发送量对照图。

由表 3 和图 5 可以看出,在预测的 45 天数据中,实际发送量与预测发送量之间最大差值为春节过后第 9 天,为 - 24 881 人;相对误差最大的为春节这一天,为 9.23 %。由此可见,实际的发送量和预测的发送量

表 3 2003 年北京路局实际直通旅客发送量与预测发送量比较

时间/天	数 据			
	实际运量/人	预测运量/人	差值/人	相对误差/%
- 19	258 957	245 645	13 312	5.14
- 18	256 803	256 225	578	0.23
- 17	254 874	262 645	- 7 771	- 3.05
- 16	254 874	262 645	- 7 771	- 3.05
- 15	266 094	264 330	1 764	0.66
- 14	260 602	271 835	- 11 233	- 4.31
- 13	268 316	262 075	6 241	2.33
- 12	271 524	259 080	12 444	4.58
- 11	250 510	259 625	- 9 115	- 3.64
- 10	269 684	245 545	24 139	8.95
- 9	242 051	255 115	- 13 064	- 5.40

续表 3 2003 年北京路局实际直通旅客
发送量与预测发送量比较

时间/天	数 据			
	实际运量/ 人	预测运量/ 人	差值/ 人	相对误差/ %
- 8	241 263	262 720	- 21 457	- 8.89
- 7	234 412	241 815	- 7 403	- 3.16
- 6	246 421	260 205	- 13 784	- 5.59
- 5	255 866	241 845	14 021	5.48
- 4	265 715	256 995	8 720	3.28
- 3	256 096	246 240	9 856	3.85
- 2	250 553	258 675	- 8 122	- 3.24
- 1	253 043	238 740	14 303	5.65
春节	248 803	225 850	22 953	9.23
1	201 833	194 250	7 583	3.76
2	146 566	141 695	4 871	3.32
3	259 840	238 495	21 345	8.21
4	298 974	298 685	289	0.10
5	304 126	295 590	8 536	2.81
6	288 686	298 980	- 10 294	- 3.57
7	341 495	336 730	4 765	1.40
8	298 062	286 465	11 597	3.89
9	274 509	299 390	- 24 881	- 9.06
10	263 372	274 285	- 10 913	- 4.14
11	266 321	257 860	8 461	3.18
12	251 595	260 640	- 9 045	- 3.60
13	250 702	258 245	- 7 543	- 3.01
14	243 224	230 240	12 984	5.34
15	267 504	257 075	10 429	3.90
16	234 887	219 895	14 992	6.38
17	296 089	278 860	17 229	5.82
18	274 483	263 400	11 083	4.04
19	260 361	265 725	- 5 364	- 2.06
20	253 727	265 670	- 11 943	- 4.71
21	265 997	262 605	3 392	1.28
22	265 653	263 920	1 733	0.65
23	264 106	258 305	5 801	2.20
24	259 570	241 785	17 785	6.85
25	220 927	212 545	8 382	3.79

5 结 论

本文在分析铁路客票数据的基础上,采用分段模糊 BP 神经网络对旅客发送量进行预测。结果表明,根据分段模糊 BP 神经网络得到的预测模型对铁路旅客发送量有很好的预测效果,说明该预测模型是可信

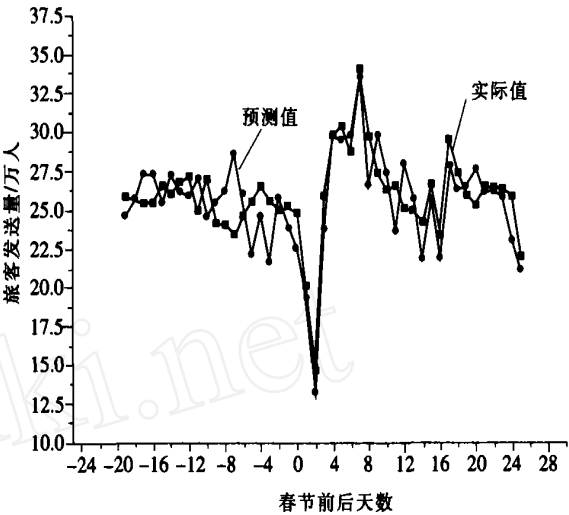


图 5 2003 年北京路局旅客实际发送量与预测量

的。
预测通常分为近期预测、中期预测和远期预测。本文引用的数据是铁路假日运输管理系统中春运期间的客票数据,是一种近期预测行为,对中期和远期预测方法的适用性有待进一步研究。另外,总运量预测只是运量预测的一个方面,在此基础上进一步考虑客流在具体发到点和线路上的分布问题,将对路网建设、投资决策和经营管理等具有更实际的参考价值。

参考文献:

[1] 刘春煌,梁明珠. 铁路客票数据仓库建设方案的研究[J]. 中国铁道科学, 2001, 22(3): 9—14.
[2] 刘春煌. 铁道部客票中心系统的设计与关键技术的实现[J]. 中国铁道科学, 2001, 22(2): 15—22.
[3] 刘胜利,曾鸣,苏宝库. 基于模糊神经网络的故障诊断及在无刷直流电机系统中的应用[J]. 中国惯性技术学报, 1999, 7(2): 49—53.
[4] 赖文彬. 面向具有空间属性系统的数据挖掘方法[D]. 北京:北京交通大学, 2004.
[5] 李洪兴,等. 工程模糊数学方法及应用[M],天津:天津科学技术出版社,1993.
[6] Hechi - Nielsen R. Theory of the back propagation neural network [J]. Proc. of IJCNN, 1989, 11: 593—603.
[7] 王艳辉. 地下动态开挖地表移动机理及智能预测系统研究[D]. 北京:北京科技大学, 2003.
[8] Ghaboussi J, Garret J H Jr, Wu X. Knowledge - Based modeling of material behavior with Neural Networks[J]. Journal of Engineering Mechanics Division, ASCE, 1991. 32—153.
[9] 闻新,周露,王丹力,熊小英. Matlab 神经网络应用设计[M]. 北京:科学出版社,2001.

(责任编辑 李淑萍 刘梅林)