

文章编号:1001-9081(2010)04-1004-04

非线性降维算法及其在医院绩效考核上的应用

李 凯¹, 黄添强^{1,2}, 余养强¹, 郭躬德¹

(1. 福建师范大学 数学与计算机科学学院, 福州 350007; 2. 清华大学 信息科学技术学院, 北京 100084)

(Lkk_Li@163.com)

摘 要:流形学习算法中的等距嵌入算法 (ISOMAP) 具有对离群点敏感的瑕疵, 针对此问题, 提出利用基于共享近邻的距离度量方式, 并充分利用了流形上对象的局部密度信息, 有效改善了算法的性能, 提高了算法的健壮性。同时, 首次尝试将该改进的流形学习算法应用于医院绩效考核。人工数据与真实数据上的实验表明, 改进的算法健壮且有效, 在绩效考核上应用成功。

关键词:非线性降维; 共享近邻; 等距嵌入算法; 离群点; 绩效考核

中图分类号: TP181 **文献标志码:** A

Nonlinear dimensionality reduction algorithm and application to hospital performance evaluation

LI Kai¹, HUANG Tian-qiang^{1,2}, YU Yang-qiang¹, GUO Gong-de¹

(1. School of Mathematics and Computer Science, Fujian Normal University, Fuzhou Fujian 350007, China;

2. School of Information Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Manifold learning algorithm ISOMAP is sensitive to the outliers. To solve this problem, the paper employed the distance measurement based on shared nearest neighbor and made a full use of the local density information of points on the manifold, which resulted in an effective improvement on the robustness of the algorithm. Meanwhile, the paper first attempted to apply the improved manifold learning algorithm to the hospital performance evaluation. The experiments on the artificial data and real-world data show that the improved algorithm is robust and effective, and the application to the performance evaluation is successful.

Key words: nonlinear dimensionality reduction; shared nearest neighbor; Isometric Mapping (ISOMAP); outlier; performance evaluation

0 引言

随着信息技术的广泛应用, 数据库中堆积了大量的多维数据, 如商业记录、气候模式、恒星光谱、人类基因分布等。在处理这些数据时, 会碰到降维问题, 尤其是非线性降维问题, 降维的目的就是找出隐藏在多维数据中有意义的低维结构, 它是人们做分类决策前的重要步骤。

流形学习算法是非线性降维的重要工具之一, 目前已有的代表性工作包括: 等距嵌入算法 (Isometric Mapping, ISOMAP)^[1]、局部线性嵌入法 (Locally Linear Embedding, LLE)^[2]、拉普拉斯特征映射 (Laplacian Eigenmaps, LE)^[3] 及局部切空间排列法 (Local Tangent Space Alignment, LTSA)^[4] 等。这类算法在处理非线性降维方面有较大优势, 但也存在内在缺陷: 算法对噪声非常敏感, 少量的噪声点将对流形拓扑结构造成极大破坏。因此, 研究健壮的流形学习算法已成为当前研究的热点和难点之一。

本文主要关注 ISOMAP 算法, 针对算法对离群点敏感的问题, 提出利用基于共享近邻的距离度量方法, 尽可能地排除了对象与离群点相连接的可能性, 保证测地线沿着流形连接, 使得原来对离群点敏感的算法更加健壮。传统的绩效考核方法有熵权法^[5]、相关系数法^[6]和主成分分析法^[7]等, 这些算法在处理非线性数据时具有明显局限性, 本文利用流形学习

算法适合处理非线性数据的特点, 将它应用于绩效考核领域, 取得了良好的效果。首先对经典 ISOMAP 算法作简要介绍, 然后分析算法的内在缺陷及问题根源, 最后提出一种改进的算法。

1 流形学习算法

1.1 ISOMAP 算法及其缺陷

ISOMAP 是基于流形学习的非线性降维算法的代表作之一。它利用测地线距离代替欧式距离, 测地线距离表现为流形上任意两点之间的最短距离, 可近似为一系列邻近点之间距离之和。因为在数据的全局几何结构未知 (通常呈非线性) 的情况下, 欧式距离只在很小的领域内才有意义。ISOMAP 通过数据间的测地线距离, 保留了数据固有的几何分布结构, 具有揭示非线性数据流形的内在特性的能力。

ISOMAP 算法在处理非线性降维方面具有明显优势, 但它也存在内在局限性: ISOMAP 算法对噪声点非常敏感, 噪声点的存在会造成流形的“短路”, 影响测地线距离的计算, 进而影响到非线性降维的效果。

噪声点之所以会造成流形的“短路”, 其本质原因如下: 如图 1 所示, X_1 与 X_2 分别为流形上的两个点, X_0 为噪声点, 它造成 X_1 与 X_2 之间的“短路”现象。根据 ISOMAP 算法原理, 流形上的每个点将与其 K 个最近邻连接构成近邻图, 当 X_1 或 X_2

收稿日期: 2009-10-15; 修回日期: 2009-12-30。

作者简介: 李凯 (1984-), 男, 福建莆田人, 硕士研究生, 主要研究方向: 数据挖掘; 黄添强 (1971-), 男, 福建莆田人, 博士, 主要研究方向: 人工智能、机器学习、数据挖掘; 余养强 (1985-), 男, 福建古田人, 硕士研究生, 主要研究方向: 数据挖掘; 郭躬德 (1965-), 男, 福建龙岩人, 博士, 主要研究方向: 人工智能、机器学习、数据挖掘。

在寻找其 K 个最近邻点时,将噪声点 X_o 当作其 K 个最近邻点之一,并连接 X_1X_o 及 X_2X_o ,这样,“ $X_1 \rightarrow X_o \rightarrow X_2$ ”这一链路被连通了;另一方面, X_o 在寻找其 K 个近邻点时,也会造成该链路的连通。这就是所谓的“短路”现象。“短路”造成流形在计算 X_1X_2 之间测地线距离时将沿“ $X_1 \rightarrow X_o \rightarrow X_2$ ”这一最短路径,而非沿流形表面,因此造成测地线距离计算不精确,进而破坏后续降维过程。因此,研究一种健壮的流形学习算法,处理现实生活中带噪声的数据,是流形学习算法得到广泛应用的前提。

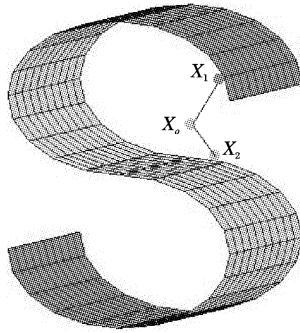


图1 S-Curve 数据集

1.2 改进的 ISOMAP

针对 ISOMAP 算法对噪声点敏感的弱点,许多学者展开了深入的研究,其中有代表性的工作包括 S-ISOMAP^[8] 和 WeightedIso^[9] 等。它们基于监督学习理论,充分利用类标号信息指导降维过程,提高算法的健壮性。此外,文献[10]提出一种基于图的网络流检测并消除离群点的算法;文献[11]提出一种基于柱状图分析的离群点检测算法等。

本文提出一种无监督的基于共享近邻距离的等距嵌入算法(Shared nearest Neighbor Distance based Isomap, SANDI),其主要思想是利用共享近邻距离,代替原始的欧式距离,降低算法对噪声点的敏感度。

共享近邻算法^[12]的定义为:在欧式空间中寻找每个点的 K 个最近邻居,然后定义两个点之间的相似性为它们共享的邻居的数目,即 K 个邻居的交集。很明显,两个点共享的邻居数目越多,它们的相似性越大。基于此概念,本文用共享近邻与 K 的比值构造一种新的距离度量方式,即共享近邻距离,如下所示:

$$D_{\text{sn}}(x_i, x_j) = 1 - \frac{SNN(x_i, x_j)}{K} \quad (1)$$

其中: $SNN(x_i, x_j)$ 表示点 x_i, x_j 共享的邻居数目, $D_{\text{sn}}(x_i, x_j)$ 为 x_i, x_j 间的共享近邻距离。

基于此距离度量,提出一种健壮的 ISOMAP 算法 SANDI,基本步骤如下:

1) 基于欧式距离计算任意两点之间的距离,找出每个点的 K 个最近邻居。

2) 基于式(1),计算任意两个点之间的共享近邻距离,并基于此距离度量重新寻找每个点的 K 个最近邻居。

3) 寻找每一个点 x_p 的最优邻居数目 K_p 。基本思想是:点 x_p 的局部邻域越稀疏,其 K_p 越小。具体过程如下:初始化 $K_p = K$,令 $Average(p)$ 等于 $\frac{1}{K_p} \sum_{i=1}^{K_p} D_{\text{sn}}(x_p, x_i)$,即表示点 x_p 与其 K_p 个最近邻居的平均距离,其中 x_i 为点 x_p 的第 i 个最近邻居。如果 $Average(p)$ 大于某一阈值 δ ($0 < \delta \leq 1$),则 $K_p = K_p - 1$,并重新计算 $Average(p)$ 的值。重复以上过程,直至 $Average(p)$ 小于阈值 δ 或者 $K_p = 1$,则此时 K_p 的值为点 x_p 的最优邻居数目。

4) 根据最优邻居数目 K_p ,淘汰邻居,基于共享近邻距离构建近邻图。连接每个点 x_p 与其 K_p 个最近邻居,并设边权重

为 $D_{\text{sn}}(x_i, x_j)$ 。需要注意的是,每个点连接不同数目的邻居。

5) 计算任意两点间的最短路径。

6) 调用经典 MDS 算法^[13],构建低维嵌入。

1.3 SANDI 算法分析

由 1.1 节对 ISOMAP 算法缺点的分析可知,ISOMAP 对噪声点的敏感表现为噪声点会使得流形发生“短路”现象,即造成链路“ $X_1 \rightarrow X_o \rightarrow X_2$ ”的连通,为此,只需保证该链路的不连通,就可防止“短路”现象的出现,降低算法对噪声点的敏感度,这也正是 SANDI 算法的基本思想。

由图 1 可知,噪声点 X_o 在两种情况下会连接点 X_1 :当 X_o 视 X_1 为其 K 个最近邻之一,或者 X_1 视 X_o 为其 K 个最近邻之一时, X_o 将连接 X_1 ,造成短路现象。因此,为了提高算法的健壮性,SANDI 算法分别从这两方面入手保证 X_o 与 X_1 之间不连通。首先,为了防止 X_o 将 X_1 视为其 K 个邻居之一,SANDI 算法根据每个点的局部密度调整图的连通状况,在步骤 2) 中,首先计算每个点与其 K_p 个最近邻居的平均距离 $Average(p)$,如果该距离大于某阈值,则说明该点的局部区域较稀疏,则减少其对应的 K_p 的值,使得该点连接较少数目的邻居。进而,由于噪声点 X_o 的局部区域较为稀疏,则其 K_p 的值较小,连接的邻居数目也较少, X_o 连接 X_1 的概率也大大降低,从而防止了“短路”现象的发生。另一方面,为了防止点 X_1 视 X_o 为其 K 个最近邻居之一,进而连接 X_1X_o ,SANDI 算法采用了基于共享近邻的距离度量方式,代替了原来的欧式距离,以提高算法的健壮性,其原理如图 2 所示: X_o 为噪声点, X_1 为流形上某一点,只需防止 X_1 视 X_o 为其 K 个最近邻居之一即可。由图 2 可知,若采用欧式距离,并设邻居数目 K 等于 7,则 X_1 的邻居为 $[X_3, X_4, X_5, X_o, X_6, X_7, X_8]$, X_o 的邻居为 $[X_{17}, X_{18}, X_{16}, X_2, X_1, X_3, X_4]$, X_o 的邻居为 $[X_{10}, X_8, X_{11}, X_7, X_{12}, X_6, X_5]$, X_1X_o 连通,造成短路。而若采用共享近邻距离,根据定义, X_1 与 X_o 共享 2 个邻居 $[X_3, X_4]$, X_1 与 X_o 共享 4 个邻居 $[X_5, X_6, X_7, X_8]$,则 $D_{\text{sn}}(X_1, X_o) = 1 - 4/6 = 0.333$, $D_{\text{sn}}(X_1, X_9) = 1 - 2/6 = 0.667$, $D_{\text{sn}}(X_1, X_9) < D_{\text{sn}}(X_1, X_o)$,此时 X_1 的邻居为 $[X_3, X_4, X_5, X_6, X_7, X_8, X_9]$, X_1 会选择 X_9 作为其邻居,而非 X_o ,从而避免了 X_1X_o 的连通。由此可知,基于共享近邻距离的算法比基于欧式距离的算法具有更强的健壮性。

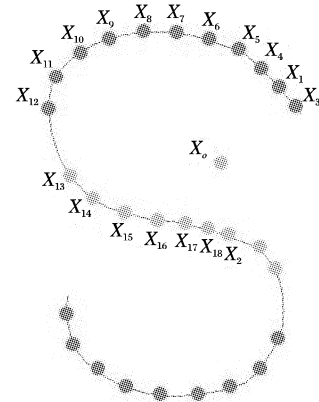


图2 SANDI 算法原理

图 2 中,设 $K = 7$,采用共享近邻距离度量,点 X_1 将选择 X_9 作为其邻居,而非噪声点 X_o 。

综上可知, SANDI 算法比原始的 ISOMAP 算法更加健壮,可以稳健地处理现实生活中许多复杂的带噪声的非线性数据。

2 实验评估

本文分别在人工数据与真实数据上进行实验,验证

SANDI 算法的有效性,并与原始 ISOMAP 算法作比较。进一步地,把 SANDI 算法应用到医院的绩效考核中去,解决现实问题。

2.1 人工数据

首先在人工数据“Swiss”数据集上进行实验,如图 3 所示,“Swiss”数据呈非线性分布,并伴随有 10 个噪声点,夹在两个平面之间的噪声点将造成流形的短路,破坏流形的拓扑结构。如图 4(a)所示,当利用 ISOMAP 构建低维嵌入时,由于噪声的存在,低维流形发生扭曲;而与此相对比的是,当利用 SANDI 算法时,流形得到很好的展开,低维嵌入较为精确,如图 4(b)所示。同时可以发现,调用 SANDI 算法时,夹在两个平面之间的噪声点将被划分到最近的一个平面上去,这是因为 SANDI 算法迫使噪声点仅连接少数的邻居点,甚至一个邻居,进而使噪声点仅与一个平面保持拓扑连通。通过在人工数据上的实验,可以直观感受到 SANDI 算法的优越性。

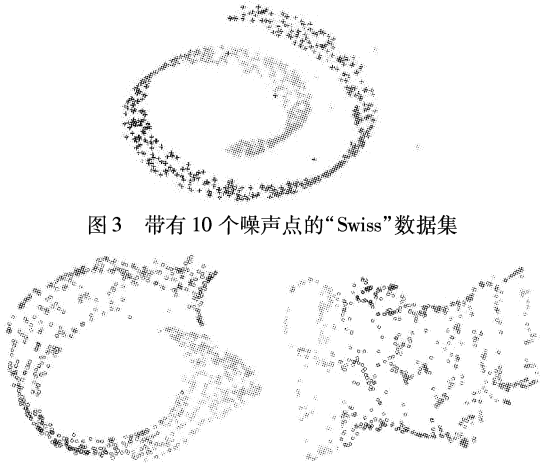


图 3 带有 10 个噪声点的“Swiss”数据集

(a) 调用ISOMAP算法的结果 (b) 调用SANDI算法的结果

图 4 对比结果

2.2 UCI 数据

接下来在几个 UCI 公共数据集^[14]上进行实验,以验证算法的有效性。为了使得实验结果具有直观可比性,将首先利用 ISOMAP 及 SANDI 算法构建低维嵌入,再调用 KNN 分类算法,最终比较两个算法的分类精确率。整个实验建立在 8 个 UCI 数据集上,采用十折交叉验证法,实验数据信息及结果如表 1、2 所示。

表 1 实验数据的基本信息

数据集	实例数	属性数	类数
Breast Cancer	569	30	2
Sonar	208	60	2
Iris	150	4	3
Heart-statlog	270	13	2
Manmagraphy	351	34	2

表 2 SANDI 与 ISOMAP 算法在分类精确率上的比较

数据集	SANDI 算法	ISOMAP 算法
Breast Cancer	0.958	0.953
Sonar	0.855	0.812
Iris	0.960	0.920
Heart-statlog	0.878	0.867
Manmagraphy	0.836	0.836
平均正确率	0.897	0.878

由表 2 可知,SANDI 算法在大多数数据集上可获得比 ISOMAP 更高的精确率,平均正确率也高于 ISOMAP 算法,体

现 SANDI 算法在处理现实中带噪声数据时的优越性,是一种健壮的非线性降维算法。

从以上在人工数据与现实 UCI 数据上的实验可以看出,SANDI 算法优于原始 ISOMAP 算法,具有较强的抗噪性。因此,SANDI 算法具有更广阔的应用空间,将其应用在医院绩效考核上具有可行性。

2.3 SANDI 算法在医院绩效考核上的应用

本节将把 SANDI 算法应用到医院绩效考核中去。表 3 列出了某医院妇科人员的基本信息,医务人员的绩效得分由 33 个考核指标计算得出,由于篇幅有限,表 3 仅列出 5 个关键指标,11 个医务人员用于分析,5 个指标均为高优指标,即值越大越优的指标。表 4 为分别利用 SANDI 算法、PCA 算法、熵权法计算得到的各医务人员的绩效得分情况。

表 3 某医院妇科人员的基本信息

姓名	门诊量	治愈好转率/%	手术次数	医疗收入/元	...	出院量
A	422	100.00	0	25 923.68	...	176
B	388	96.00	36	252 225.24	...	2
C	108	100.00	5	15 652.51	...	55
D	326	96.55	50	417 794.55	...	113
E	26	0	0	1 924.11	...	0
F	8	0	0	901.15	...	0
G	727	100.00	8	264 804.64	...	62
H	139	100.00	15	203 477.40	...	63
I	124	100.00	1	202 244.38	...	56
J	233	100.00	1	255 723.13	...	75
K	2	0	0	472.54	...	0
...

表 4 几种算法计算出的医务人员的绩效得分情况算法

姓名	绩效得分		
	SANDI 算法	PCA 算法	熵权法
A	0.505	0.289	0.361
B	0.870	0.777	0.697
C	0.570	0.332	0.524
D	1.000	1.000	1.000
E	0.100	0.106	0.105
F	0.268	0.123	0.115
G	0.974	0.769	0.727
H	0.870	0.559	0.569
I	0.707	0.536	0.570
J	0.567	0.595	0.595
K	0.281	0.100	0.100
...

已知领域专家对该医院医务人员进行考核,得到的绩效考核排名为:D>G>B>H>I>J>C>A>F>E>K>…。将各算法与专家排名进行对照,相对正确率如表 5。

表 5 各算法与专家排名正确率

算法	正确率/%
PCA	81.80
熵权法	80.70
SANDI	90.90

其中,SANDI 算法排名正确率为:90.9%,而 PCA 算法排名正确率为:81.8%,熵权法排名正确率为:80.7%。因此 SANDI 算法可得到比传统算法更为精确的绩效排名,实用性较强。

从表5可以看出,SANDI算法在医院绩效考核上的应用是成功的,可以较准确地反映出各评价对象的业绩水平,作为绩效考核的有力工具。非线性降维算法在绩效考核中的应用,是前人未曾涉及的成功尝试。

3 结语

本文提出一种基于共享近邻的健壮的非线性降维算法SANDI,解决了传统流形学习算法对噪声点敏感的问题,在人工数据及现实UCI数据上的实验表明,SANDI算法可以有效降低对噪声点的敏感度,提高算法的健壮性。同时,将SANDI算法应用于医院绩效考核系统中,尝试将该健壮的非线性降维算法应用于绩效考核,具有更大的应用空间。

参考文献:

- [1] TENENBAUM J B, SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. *Science*, 2000, 290(5500): 2319-2323.
- [2] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by local linear embedding [J]. *Science*, 2000, 290(5500): 2323-2326.
- [3] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering [C]// *Proceedings of Neural Information Processing Systems Conference*. Cambridge, MA: MIT Press, 2001: 585-591.
- [4] ZHANG Z Y, ZHA H Y. Principal manifolds and nonlinear dimension reduction via local tangent space alignment [J]. *SIAM Journal of Scientific Computing*, 2004, 26(1): 313-338.
- [5] 申富平, 马玉芳, 戎辉辉. 中国旅行社企业财务绩效评价指标研

究——基于熵值法的视角[J]. *经济与管理*, 2008, 22(6): 81-84.

- [6] 赵亚丽, 西广成, 刘艳骄. 用相关系数法探讨齿痕舌出现的特点[J]. *中国中医基础医学杂志*, 2003, 9(9): 77-79.
- [7] 朱颖超. 基于主成分分析法的炼油企业绩效评价[J]. *工业技术经济*, 2008, 27(6): 81-84.
- [8] GENG X, ZHAN D C, ZHOU Z H. Supervised nonlinear dimensionality reduction for visualization and classification [J]. *IEEE Transactions on Systems*, 2005, 35(6): 1098-1107.
- [9] VLACHOS M, DOMENICONI C, GUNOPULOS D, *et al.* Non-linear dimensionality reduction techniques for classification and visualization [C]// *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM Press, 2002: 645-651.
- [10] CHOI H, CHOI S. Robust kernel isomap [J]. *Pattern Recognition*, 2007, 40(3): 853-862.
- [11] CHEN H F, JIANG G F, YOSHIHIRA K J. Robust nonlinear dimensionality reduction for manifold learning [C]// *18th International Conference on Pattern Recognition*, Washington, DC: IEEE Press, 2006: 447-450.
- [12] JARVIS R A, PATRICK E A. Clustering using a similarity measure based on shared nearest neighbors [J]. *IEEE Transactions on Computers*, 1973, C-22(11): 1025-1034.
- [13] COX T, COX M. *Multidimensional scaling* [M]. London: Chapman & Hall, 1994.
- [14] UCI repository of machine learning databases[DB/OL]. [2008-10-12]. <http://www.ics.uc.edu/~mllearn/MLRepository.html>.

(上接第1003页)

CA1419:D5(8:14:0)→E7(8:14:16)→E8(8:14:29)→F6(8:14:45)→F5(8:14:58)→F4(8:15:37)→F3(8:16:22)→F2(8:17:7)→F1(8:17:52)→F0(8:18:15)→I0(8:18:38)。

MU5102:C5(8:14:0)→B4(8:14:16)→B3(8:14:31)→A3(8:14:47)→A2(8:15:22)→A1(8:15:57)→A0(8:16:21)→R0(8:16:55)。

4 结语

本文应用改进蚁群协同算法和滑动控制窗口相结合的方法对滑行调度进行了研究,保证了滑行调度零冲突和滑行调度合理化。约束条件的设计保证了滑行调度的零冲突,保证了滑行调度的安全可靠。目标函数的设计考虑了滑行道的滑行时间和在停机位的等待时间。本文提出的滑行窗口控制的办法考虑了不同时间段间航班滑行路径的相互影响。改进蚁群协同算法使多个航班可以并行处理,很好地对多个航班的滑行路径进行了协同。针对国内某枢纽机场8:00~8:30的航班进行滑行调度仿真,并与普通蚁群协同算法比较,改进蚁群协同算法比普通蚁群协同算法有明显的优势。国内对滑行调度研究较少处于起步阶段,有很大的研究空间,本文对滑行道的研究考虑离港航班的优先级高于进港航班的优先级,没对不同机型的飞机滑行优先级做过多考虑,可作为今后改进研究的方向。

参考文献:

- [1] 刘兆明, 葛宏伟, 钱峰. 基于遗传算法的机场调度优化算法[J]. *华东理工大学学报:自然科学版*, 2008, 34(3): 392-398.
- [2] PESIC B, DURAND N, ALLIOT J M. Aircraft ground traffic op-

timisation using a genetic algorithm[C]// *GECCO'01: Proceedings of the Genetic and Evolutionary Computation Conference*. San Francisco: Morgan Kaufmann, 2001: 1397-1404.

- [3] 孙芸. 民用机场相关决策问题研究[D]. 南京: 南京航空航天大学, 2008.
- [4] ATKINS S, JUNG Y, BRINTON C, *et al.* Surface management system field trial results[EB/OL]. [2009-07-01]. http://aeronautics.arc.nasa.gov/assets/pdf/atkins_09_04.pdf.
- [5] CHENG V H L, ALTO P. Airport surface operation collaborative automation concept[EB/OL]. [2009-07-01]. http://pdf.aiaa.org/preview/CDReadyMGNC03_664/PV2003_5773.pdf.
- [6] KEITH G, RICHARDS A. Optimization of taxiway routing and runway scheduling[EB/OL]. [2009-07-01]. http://pdf.aiaa.org/preview/CDReadyMGNC08_1853/PV2008_6827.pdf.
- [7] 陈世林, 胡明华, 张洪海. SMS中基于冲突探测的滑行轨迹预测算法研究[J]. *四川大学学报:自然科学版*, 2008, 45(6): 1357-1361.
- [8] VISSER H G, ROLLING P C. Optimal airport surface traffic planning using mixed-integer linear programming[EB/OL]. [2009-07-01]. http://pdf.aiaa.org/preview/CDReadyMATIO03_832/PV2003_6797.pdf.
- [9] WOOD Z, KISTLER M, RATHINAM S, *et al.* A simulator for modeling aircraft surface operations at airports[C]// *AIAA Modeling and Simulation Technologies Conference*. Chicago: Cussian Associate Inc, 2009: 162-172.
- [10] 黄席樾, 向长城, 殷礼胜. 现代智能算法理论及应用[M]. 北京: 科学出版社, 2005.
- [11] 李永先. 车辆路径问题的仿真模型及优化方法研究[D]. 大连: 大连理工大学, 2007.