基于潜在类别模型的航空旅客分类

顾兆军,王 伟,李晓红
(中国民航大学 计算机学院,天津 300300)

摘 要:潜在类别模型是用潜在的类别变量来解释外显的类别变量之间的关联,使外显变量之间的关系通过潜在类别变量来估计,进而维持其局部独立性。为研究民航旅客的选择行为偏好、改进航空公司收益管理策略,文中引进潜在类别模型,然后从PNR数据中选择合适的观察变量,概率参数化后带入模型进行建模,利用Mplus软件对模型进行求解评价,最终得到了最合理的民航旅客分类。由于是基于订票数据的,所以与以往研究相比,此方法从根本上避免了因可靠性带来的偏差风险。

关键词:选择行为;潜在类别概率;条件概率;潜在聚类分析
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2012)04-0182-05

Classification of Airline Passengers Based on Latent Class Model

GU Zhao-jun, WANG Wei, LI Xiao-hong
(College of Computer, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Latent class model uses the categorical latent variables to explain the connection between categorical manifest variables, makes the relationship between categorical manifest variables estimated by the categorical latent variables, and maintains its local independence. In order to study the airline passengers' choice behavior and improve strategy of revenue management of the airline, the latent class model for the PNR data is used to set up a model, and the Mplus software is also used to solve and evaluate, finally the best and reasonable airline passengers' classification is obtained. Compared with the previous researches, this method fundamentally avoids the risk of response bias.

Key words: choice behavior; latent class probability; conditional probability; potential cluster analysis

0 引 言

航空旅客选择行为是一种旅客对航空服务的消费和购买决策,在这个过程中,旅行目的、舱位等级、离港时刻、提前订座时间等诸多因素都影响着旅客最终的购买行为。这些因素使得航空旅客具有不同的类型和不同的需求,不同的需求又产生不同的选择偏好,如:舱位等级偏好、离港时刻偏好等等。旅客面临众多不同票价产品的选择时,都会选择具有最大效用的产品。与一般的商品相比,机票具有不可触摸、不可事先体验、服务的消费与生产同步等特点,使旅客购买机票的选择行为更加复杂[1,2]。

分析研究民航旅客的选择行为偏好对于制定一系列航线规划策略:网络结构设计、时刻表制定、定价和收益管理等都是至关重要的,有助于航空公司进一步了解旅客选择特征和需求、了解市场并赢得市场。航空公司需要一个更好的模型来对新的策略进行分析支持,这些策略包括以座位选择权和改票退票政策为划分的一系列品牌套餐计划。航空旅客分类是分析理解航空旅客选择行为的关键步骤。合理分类后,可以对旅客分类特征分析、旅客对航空公司偏好分析、旅客购买行为分析及票价影响分析等进行深入的研究。文中首先对航空旅客的类型进行合理的划分,方便之后对旅客选择偏好和行为特征的研究,并提出航空公司收益管理策略。

1 潜在类别模型的基本原理

社会科学研究所关心的研究议题往往众多、抽象而且无法直接观察,为了进行测量与分析,往往使用一些间接测量的方法,获得片面的信息,然后利用潜在变量的概念来整合这些间接测量信息,进而掌握抽象特质的状态。

潜在类别模型(latent class modeling; LCM)是讨论潜在变量(latent variable)的模型化分析技术。它与传统的因素分析最大的不同在于变量的形式:因素分析

收稿日期:2011-08-27;**修回日期**:2011-11-29
基金项目:中国民航局科研基金项目(MHRD200808)
作者简介:顾兆军(1966-),男,山东蓬莱人,教授,研究方向为计算机网络与信息安全、搜索引擎、民航信息系统;王 伟(1986-),男,山东滨州人,硕士研究生,研究方向为网络与信息安全。

处理连续变量;LCM 可以处理类别变量。类别变量虽然不像连续变量具有丰富的变异与可计数单位来进行多样化的统计分析,但是类别变量却拥有容易取得、容易操作等特点。因此潜在类别模型是研究纷杂的类别信息时有力的分析工具。更重要的是,潜在类别模型把类别信息与潜在变量的观念加以结合,在具有方法学价值的同时还拥有巨大的实际意义。一个完整的潜在类别模型的建立需要经过模型的概率参数化、参数的估计、模型识别、拟合优度评价、潜在聚类分析与结果解释等几个步骤[3~6]。

1.1　概率参数化

概率参数化(probabilistic parameterization)是 LCM 最根本的原理,它将类别变量的概率转换成参数的形式代入模型[3]。传统的 LCM 参数涉及两种类别变量:可以观察与测量的外显变量(manifest variable)与不可观察的潜在变量(latent variable),还有两种不同的概率值:潜在类别概率(latent class probabilities)和条件概率(conditional probabilities)。在潜在类别模型中,假定任意两个外显变量之间的关联可以由潜在变量来解释和估计。下面是具有 A、B、C 三个外显变量的潜在类别模型的公式:

$$\pi_{ijk}^{ABC} = \sum_{t=1}^{T} \pi_t^X \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X} \pi_{kt}^{\bar{C}X} \tag{1}$$

公式(1)中,π_{ijk}^{ABC} 表示一个潜在类别模型的联合概率(joint probability,为各潜在类别概率的总和)。π_t^X 表示观测数据归属于某一个潜在变量 X 的特定潜在类别的概率,亦即 $P(X = t)$,$t = 1,2,\cdots,T$。各潜在类别的相对大小(概率)代表它在潜在变量中具有的地位。所有潜在类别的概率总和为 1:$\sum_{t=1} \pi_t^X = 1$ 。$\pi_{it}^{\bar{A}X}$ 则表示属于第 t 个潜在类别的受测者对于观察变量 A 的水平 i 作出反应的概率,亦即 $P(A = i \mid X = t)$,$i = 1,2,\cdots, I$,以此类推。根据概率和条件概率的性质,有如下条件成立:

$$\sum_i \pi_{it}^{\bar{A}X} = \sum_j \pi_{jt}^{\bar{B}X} = \sum_k \pi_{kt}^{\bar{C}X} = 1 \tag{2}$$

1.2　模型估计与模型识别

当提出一个假设模型并进行概率参数化,建立模型之后,要计算参数的最终解,并且要注意参数估计时的识别问题,否则模型的计算是错误的。

在潜在类别模型中,模型求解所用到的最主要的方法是极大似然法,在迭代过程中所使用的算法有 EM(Expectation Maximization)、NR(Newton Raphson)等不同算法,其中 EM 算法是 LCM 中最常使用的;EM 算法不受初始值选择影响的稳健性,但是迭代次数比较多,并且不提供标准误差的估计值,而 NR 算法正好相反;有研究者建议在估计初期先以 EM 算法进行迭代,当接近收敛时,可应用 NR 算法继续计算,如此将可兼顾 EM 的稳定性与 NR 的速度优点[3]。

LCM 模型中的参数要得到一组最佳解,必须进行模型的识别判断,这就要求参数的数目必须小于自由度。自由度要大于 0,否则将造成模型不能识别的后果。自由度大于 0 也不一定能让模型具有可识别性。遇到模型无法识别的情况,可以将部分参数设定限制,改变概率估计的公式,提高模型估计的数学条件。

1.3　模型的评价

为了找出更简洁,具有较少参数,又具有较好拟合优度的模型,还需要对模型进行评价。在模型评价中已经得到广泛使用的有 4 种指标:Pearsonχ^2、似然比 χ^2、Akaike 信息准则(AIC)、Bayesian 信息准则(BIC)[5]。其中利用 AIC 指标作为模型适配标准时,倾向于选择出较简效、单纯的模型,比较适用于待估参数较少、自由度较大的模型。由于 AIC 指标并没有考虑样本数的影响,因此当样本数越大时,AIC 概率缺乏推导的渐进性,此时可采用 BIC 指标。一般会将四种主要指标结合起来进行模型好坏的评价。

1.4　潜在聚类分析

分类是模型的最终目的。潜在聚类分析是创造一个新的类别变量,即潜在类别变量,来分析观察值的后验类别属性(posterior membership),将所有的观察值分类到适当的潜在类别当中。分类概率的计算过程:

$$\pi_{ijkt}^{\bar{X}ABC} = \frac{\pi_{ijkt}^{ABCX}}{\sum_t^T \pi_{ijkt}^{ABCX}} \tag{3}$$

分类的基本原理是贝叶斯理论,利用式(3)求出潜变量 X 的条件概率 $\pi_{ijkt}^{\bar{X}ABC}$ 后,根据 $\pi_{ijkt}^{\bar{X}ABC}$ 的大小判断某一条目到底属于哪一类。如果某一类的潜在类别概率最大,则此个体就归为该类。

1.5　潜在类别模型操作形式

潜在变量模型受到重视的另一个重要理由是社会科学所探究的课题是复杂的人类行为经验与社会现象,研究者所搜集到的资料往往是片段、交错复杂的数据,这些原始数据必须进一步加以整理,化简成为清楚明确的研究变量,以进行后续的统计分析,进而能够对于现象进行解释。1970 年代以来盛行的主成分分析与因素分析,乃至于当代流行的结构方程模式(structural equation modeling, SEM),最重要的功能就是可以协助研究者将一堆测量数据整并、萃取出少数的几个主成分(component)或因素(factors)来进行精简、有效率的分析。潜在变量模式不仅可以协助研究者进行抽象概念的研究,它的另一种功能亦能协助进行资料化简与整并等这种不需要理论基础,纯粹是一种探索性的资料处理作业。简言之,LCM 不仅可以从事资料化

简的探索性(exploratory)作业,也可以配合研究者的学理内涵与理论需求,进行验证性(confirmatory)研究。基于研究目的的不同,潜在类别模型可以区分为探索性与验证性两种不同的操作形式。

由于缺乏民航旅客分类参数设定的经验,文中主要采用的是探索性潜在类别模型分析,下面是探索性模型分析的过程:

1. 对只有一个类别的初始模型进行估计;

2. 逐步增加潜在类别的个数,分别求解估计模型;

3. 对已建立的几个模型进行适配性检验与差异检验,选出最佳模型;

4. 结合外显变量的特性,进行类别的命名与参数估计结果整理;

5. 潜在聚类分析,决定各观察值的归属类别。

2 航空旅客选择行为的潜在类别模型

以往对航空旅客的分类主要有两种:第一种,调查旅客的出行目的,将他们分为商务旅客和休闲旅客(包括旅游、探亲及其他类型旅客),这两类旅客群体具有大不相同的自身特征和选择的偏好。这种划分是航空运输业中一种标准式的分类。第二种,对传统的旅客划分进行市场细分,将商务旅客划分为紧急商务旅客和计划商务旅客;将休闲旅客划分为紧急休闲旅客和计划休闲旅客[2]。

传统的航空旅客选择行为研究和分类大都是基于对航空旅客的调查问卷得来的数据,收集消费者细节信息(如收入、年龄、性别、旅行方式),调查旅行性质(出行目的)等[7~10],这种数据提供了很大的便利。然而调查问卷的可靠性使得这些研究都有着偏差的风险,并且数据在设计上无法充分体现航空市场上产品结构的复杂性,以及定价和收益管理策略对旅客选择的影响。国外已经有学者利用订票数据来研究民航旅客的选择行为[11],国内相应的研究则很少。文中提出了基于订票数据的选择行为模型,这样就没有了可靠性偏差,提出一个更直观化的市场划分。

2.1 PNR 数据

下面,利用已有的 PNR 数据进行航空旅客的分类。PNR 是航空旅客订座记录,即英文 Passenger Name Record 的缩写,包含众多字段:BOOK_ID(记录编号)、BOOK_PNR_TIME(订座时间)、BOOK_FLT_DPT_TIME(起飞时间)等等。它反映了旅客的航程、航班座位占用的数量及旅客信息。

2.2 选取分类属性

订座信息中不记录出行目的,但可以从订座信息获得其他有关旅行性质和旅客性质的信息,这些信息从侧面和局部反映了旅客的出行目的。用相似的方法,可以借由潜在类别模型分析来还原出行目的。这些都有赖潜在变量模型的应用。

在接下来的实验中,对现有 PNR 数据进行数据预处理后选择出以下分类属性(见表 1):

表 1 类别属性

项目	否(0)		是(1)	
	N	%	N	%
A:是否 VIP	30465	99.5	141	0.50
B:是否团队购票	28930	94.5	1676	5.50
C:是否经济舱	6740	22.0	23866	78.0
D:离港-订座日期差是否<=3	16351	53.4	14255	46.6
E:离港时刻是否上午	18762	61.3	11844	38.7
F:离港日期是否周末	22660	74.0	7946	26.0

旅客的性质(是否 VIP)、购票的方式(是否团队购票)、舱位性质(是否经济舱)、离港日期与订票日期之间的天数、离港时刻,以及离港日期是否是节假日等都反映着旅客的出行目的。把这些属性的概率参数化,代入模型。

2.3 模型求解方法

可以为潜在类别模型进行建模的软件有很多:最著名的当属 Statistical Innovation 公司(Vermunt and Magdison,2005)专门为潜在分类选择模型建模而设计的软件 Latent Gold Choice[12]。由 Bengt Muthen 与 Linda Muthen 开发的 Mplus[13] 也被广泛应用于 LCM 建模。著名统计软件 SAS 9.0 可以利用外挂模块 PROC LCA 与 PROC LTA 进行潜在类别分析。下面利用 Mplus 软件对模型进行求解。

3 实验结果

3.1 模型评价与选择

为了选择合适的潜在类别模型,采用探索性潜在类别分析,即从潜在类别数目为 1 的初始模型开始,一共拟合了 9 个潜在类别模型。表 2 给出了 9 个潜在类别模型的拟合结果:

表 2 分类评价表

Class	HO	FP	AIC	BIC	X^2	G^2
1	-82627.274	6	165266.548	165316.522	4273.634	3990.560
2	-81380.251	13	162786.502	162894.778	1648.092	1496.513
3	-80864.442	20	161768.883	161935.462	906.912	464.895
4	-80780.948	27	161615.896	161840.778	960.342	297.908
5	-80770.445	34	161608.889	161892.074	333.009	276.900
6	-80744.877	41	161571.755	161913.242	836.764	225.766
7	-80695.322	48	161486.644	161886.434	142.832	126.655
8	-80668.690	55	161447.380	161905.472	67.559	73.391
9	-80664.995	62	161453.990	161970.385	54.483	66.001

由拟合信息可知:当潜在类别数目为 3 时,模型满足了数据拟合的要求(AIC = 161768.883, BIC = 165316.522, χ^2 = 906.912, G^2 = 464.895, $P > 0.05$),

并且此时的 BIC 接近最小,χ^2 与 G^2 也比类别数目为 2 时小很多,类别数目继续增加时参数数目明显增多,模型拟合优度未见明显改善。故可以选择包含 3 个潜在类别数目的模型作为分析的理想模型。

3.2 模型参数估计

对类别数目为 3 的模型,利用 EM 算法对潜在类别概率和潜在类别下各项目条件概率的估计,结果如表 3:

表 3　潜在类别条件概率与潜在类别概率估计

项目		潜在类别		
		Cluster 1	Cluster 2	Cluster 3
A	1	0.018	0.000	0.001
	0	0.982	1.000	0.999
B	1	0.009	0.000	0.130
	0	0.991	1.000	0.870
C	1	0.060	1.000	1.000
	0	0.940	0.000	0.000
D	1	0.277	1.000	0.094
	0	0.723	0.000	0.906
E	1	0.373	0.370	0.410
	0	0.627	0.630	0.590
F	1	0.237	0.244	0.287
	0	0.763	0.756	0.713
潜在类别概率		0.22041	0.39959	0.37999

表 3 中的概率值是属于第 t 个潜在类别的受测者对于观察变量 A 的水平 i 作出反应的概率,亦即 $P(K=i \mid X=t)$,K 的取值集合是 $\{A,B,C,D,E,F\}$,i 的取值范围是 $\{0,1\}$,t 为 $\{1,2,3\}$。根据概率和条件概率的性质,有如下条件成立:

$$\sum_i \pi_{it}^{\bar{A}X} = \sum_i \pi_{it}^{\bar{B}X} = \sum_i \pi_{it}^{\bar{C}X} = \sum_i \pi_{it}^{\bar{D}X} = \sum_i \pi_{it}^{\bar{E}X} = \sum_i \pi_{it}^{\bar{F}X} = 1 \tag{4}$$

在观察值属于某一确定类的条件下,属性取值为 1 的条件概率图如图 1 所示:

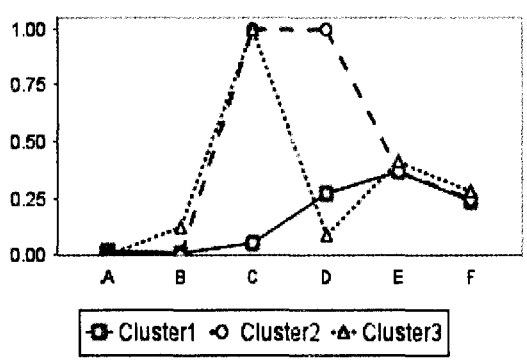

图 1　确定类的条件概率

从图 1 的折线图可以看到,在属性 A、B 处,由于取值为 1 的概率很小,导致 3 种类别条件概率的差异显示不出来,在属性 C、D 处,三者差异很大:类别 2、3 中所有的旅客选择均为经济舱,只有类别 1 的旅客选择了非经济舱;类别 2 的旅客最趋向于在离港三天之内订票。在属性 E、F 处三者无明显差异。

接下来计算属性值确定的条件下观察值属于某一类的概率,如属性 A 取值为 1 的条件下,观察值属于第一类的概率计算过程如下:

$$P(X=t \mid A=1) = \pi_{11}^{\bar{X}A} = \frac{\pi_{11}^{AX} \times \pi_1^X}{\pi_1^A} = \frac{0.018 \times 0.22041}{0.005} = 0.793476 \tag{5}$$

式(5)表示订票信息中属于 VIP 的记录,被分到 cluster1 的概率是 79.3476%。如此计算可得到图 2。图中 $A1$ 表示属性 A 取值为 1,$A0$ 表示 A 取值为 0,以此类推。

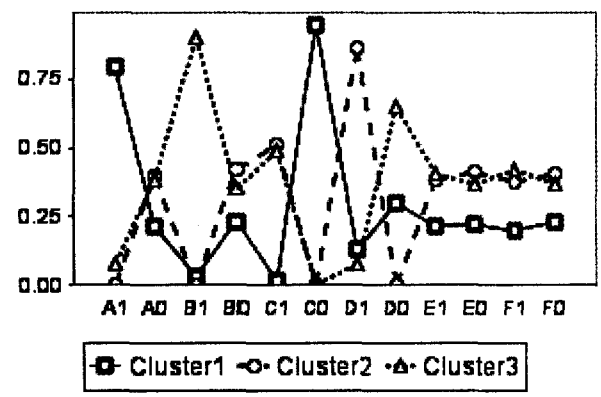

图 2　确定属性值的条件概率值

3.3 聚类分析

由于将每一记录划分到不同的类别中,是根据概率的结果,它并不是完全准确的,在分类过程中可能会出现一些错误。Clogg(1979,1981)提出了两种用来计算归类正确率的方法。

(1) 正确分类百分比,算法如下:

$$100 \times \sum_{ijk} (\pi_{tijk}^{\bar{X}ABC} \times p_{ijk}) \tag{6}$$

其中,p_{ijk} 表示 $\{i,j,k\}$ 组合占总记录数目的百分比。

(2) lambda(λ)(Goodman & Kruskal,1954),算法如下:

$$\lambda = \frac{E_1 - E_2}{E_1} \tag{7}$$

其中,$E_1 = 1 - \pi_t^X$,表示将记录分入潜在类别概率最大的类别时所引起的错误率,π_t^X 表示 T 个潜在类别中最大的潜在类别概率。$E_2 = \sum_{ijk} (1 - \pi_{tijk}^{\bar{X}ABC}) p_{ijk}$。

用第一种方法,经过计算可得正确分类的概率为 96.93%,说明分类基本上是正确的。表 4 是各组合的分类结果(由于组合比较多,一共 64 种,只列举了其中几个)。

结合图 1、图 2 以及具体的分类结果,可以总结三

个类别旅客的特点并分别命名：

表 4 分类结果表

A	B	C	D	E	F	数目	Cluster
0	0	0	0	0	0	2291	1
0	0	0	0	0	1	681	1
0	0	1	0	0	0	4139	3
0	0	1	1	0	0	5703	2
0	1	1	0	0	0	764	3
1	0	0	1	0	0	28	1
1	0	1	0	0	1	8	3

Cluster1：VIP 旅客几乎全部属于此类，此类旅客大部分不是团队购票，且选择的是非经济舱。将其命名为：完全公务型旅客；

Cluster2：旅客大部分不是团队购票，选择经济舱，且在离港前三天以内订票。将其命名为：临时混合型旅客；

Cluster3：团队购票，选择经济舱，离港前三天以外订票。将其命名为：计划休闲型旅客。

4 结束语

文中将潜在类别模型应用到航空领域，对旅客进行了分类，并得出合理的解释。为以后研究航空旅客选择行为做好铺垫。由于订票数据比较难以获取，文中实验所用数据属性仍然不够丰富，在接下来的研究中希望能够获取更加准确丰富的订座信息，以研究民航旅客的选择行为分类及偏好。

参考文献：

[1] 陈 剑，肖勇波，刘晓玲，等. 基于乘客选择行为的航空机票控制模型研究[J]. 系统工程理论实践，2006，26(1)：65-75.

[2] 梅 虎，朱金福，汪 侠. 我国航空旅客航班舱位选择行为研究[J]. 经济问题探索，2006(12)：89-93.

[3] 邱皓政. 潜在类别模型的原理与技术[M]. 北京：教育科学出版社，2008.

[4] Hagenaars J A，McCutheon A L. Applied latent class analysis [M]. Cambridge：Cambridge University Press，2002.

[5] 郭小玲，裴磊磊，张岩波. 潜在类别模型及数据模拟分析[J]. 数理医药学杂志，2009，22(6)：631-635.

[6] 裴磊磊，郭小玲，张岩波，等. 抑郁症患者单核苷酸多态性(SNPs)分布特征的潜在类别分析[J]. 中国卫生统计，2010(1)：7-10.

[7] 王春兰. 航空公司收益管理中旅客舱位选择行为研究[D]. 南京：南京航空航天大学，2006.

[8] 梅 虎. 航空旅客选择行为及其在收益管理中的应用研究[D]. 南京：南京航空航天大学，2007.

[9] 王 爽，赵 鹏. 基于 Logit 模型的客运专线旅客选择行为分析[J]. 铁道学报，2009(3)：6-10.

[10] 梅 虎，朱金福，汪 侠. 旅客航班选择模型研究：变精度粗集方法[J]. 管理评论，2007(3)：27-32.

[11] Carrier E. Modeling the Choice of an Airline Itinerary and Fare Product Using Booking and Seat Availability Data[D]. Massachusetts：Massachusetts Institute of Technology，2008.

[12] Vermunt J K，Magdison J. Technical Guide for Latent Gold Choice 4.0：Basic and Advanced[M]. Belmont，MA：Statistical Innovations Inc.，2005.

[13] Muthen L K，Muthen B O. Mplus Statistical Analysis With Latent Variable User's Guide[M]. 5th ed. [s. l.]：[s. n.]，2008.

(上接第 181 页)

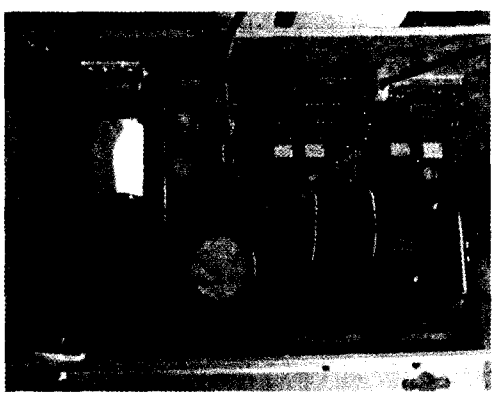

图 5 电源实物图

参考文献：

[1] 李俊峰，王斯成. 中国光伏发展报告[M]. 北京：中国环境科学出版社，2007.

[2] Chen Wei，Shen Hui，Deng Youjun，et al. Application and Development of the Inverter in the Photovoltaic System[J]. Power Electonics，2006，40(4)：130-133.

[3] 冯垛生. 太阳能发电原理与应用[M]. 北京：人民邮电出版社，2007.

[4] 陈道炼. DC/AC 逆变技术及其应用[M]. 北京：机械工业出版社，2003.

[5] 马尼克搭拉. 精通开关电源设计[M]. 北京：人民邮电出版社，2008.

[6] Pressman A I. Switching Power Supply Design[M]. 2nd ed. Beijing：Publishing House of Electronics Industry，2005.

[7] 比林斯. 开关电源手册[M]. 北京：人民邮电出版社，2006.

[8] 刘君凤. 现代逆变技术及应用[M]. 北京：电子工业出版社，2006.

[9] 曹远跃. 适用于独立光伏系统的正弦波逆变电源研制[J]. 机电技术，2008，31(4)：69-72.

[10] 范玲莉，邓 焰. 基于 LM25037 的车载便携式 SPWM 逆变器设计[J]. 电子技术应用，2009，35(7)：11-14.

[11] Liu Shulin，Liu Jian，Chen Yongbing. Output Ripple Voltage of Boost Converter and Design of Its Minimal Inductance[J]. Journal of Xi'an Jiaotong University，2007，41(6)：709-716.

[12] 张 跃，马永刚，李自立. 家用小型光伏电源系统的设计及应用[J]. 太阳能，2007(1)：26-28.