文章编号: 1001-1986(2015)06-0070-05

基于粗糙集的 Logistic 回归模型在矿井突水模式识别中的应用

王江荣[1]，黄建华[1]，罗资琴[2]，文 晖[1]

(1. 兰州石化职业技术学院信息处理与控制工程系，甘肃 兰州 730060;
2. 兰州石化职业技术学院石油化学工程系，甘肃 兰州 730060)

摘要: 矿井突水模式识别是一个非正态、非线性和高维数据处理问题，也是二分类问题。使用粗糙集属性约简算法对样本数据降维，建立 Logistic 回归模型，并利用粒子群算法对模型参数优化。该模型对建模样本突水模式识别正确率为 90%，对测试样本突水模式识别正确率为 100%，效果好于数据不降维的 Logistic 回归模型。该模型克服了线性回归分析解决二分类问题存在的不足，为矿井突水模式识别提供了一种新思路、新方法。

关 键 词: 矿井突水；模式识别；粗糙集属性约简；Logistic 回归模型；粒子群算法

中图分类号: P641.4；TD741 **文献标识码:** A DOI: 10.3969/j.issn.1001-1986.2015.06.014

Application of Logistic regression model based on rough set in recognition of mine water inrush pattern

WANG Jiangrong[1], HUANG Jianhua[1], LUO Ziqin[2], WEN Hui[1]

(1. *Department of Information Processing and Control Engineering*, *Lanzhou Petrochemical College of Vocational Technology*, *Lanzhou* 730060, *China*; 2. *Department of Petroleum Chemical Engineering*, *Lanzhou Petrochemical College of Vocational Technology*, *Lanzhou* 730060, *China*)

Abstract: The mine water bursting pattern recognition is a non normal, nonlinear and high dimensional data processing problem, but also a binary-class problem. The attribute reduction algorithm of rough set was used to reduce the dimension of the sample data, to establish Logistic regression model, and particle swarm algorithm was used to optimize model parameters. The recognition accuracy of the model was 90% for water inrush mode of the modeling samples and 100% for water inrush mode of the testing samples, the effect was better than that of the Logistic regression model without dimensionality reduction. The model overcomes the shortcomings of the linear regression analysis for the solution of the binary-class problem, provides a new method for pattern recognition of mine water inrush.

Key words: mine water inrush; pattern recognition; rough set attribute reduction; Logistic regression model; particle swarm algorithm

矿井突水严重危害着煤炭的开采、矿工的生命和财产安全，因此，对矿井突水的预测预报工作十分必要[1-2]。影响煤矿突水的主要因素(指标或特征)有：含水性、水压、隔水层厚度、突水系数、导水性、构造发育、岩性组合、推进步距和工作面斜长等 9 个因素[3]。而矿井突水通常与其特征表现之间不存在明确的对应关系，多种特征之间还存在着复杂的耦合关系等，它们与矿井突水之间也存在着非线性关系，这些特点给矿井突水模式识别工作带来较大难度。另外，矿井突水是一种二分类变量(取 1 或 0 两种值，其中 1 表示发生了突水，0 表示未发生突水)，我们采用 Logistic 回归模型对这种二分类变量进行回归分析，该回归模型实际上是一种非线性分类统计方法[4]。建模时首先使用粗糙集理论中的属性约简算法对采集到的矿井突水历史样本数据进行约简，剔除与

收稿日期: 2014-07-03
基金项目: 甘肃省科技厅项目(1204GKCA004)；甘肃省财政厅专项资金立项资助(甘财教[2013]116)
作者简介: 王江荣(1966—)，男，甘肃静宁人，教授，从事智能算法、数值计算、综合评判技术等方面的研究. E-mail：lzshwjr@163.com
引用格式: 王江荣，黄建华，罗资琴，等. 基于粗糙集的 Logistic 回归模型在矿井突水模式识别中的应用[J]. 煤田地质与勘探，2015，43(6)：70–74.

决策信息不相关的属性(属性并不是同等重要的，还存在冗余，这不利于做出正确而简洁的决策)，同时去除冗余建模样本数据，然后利用约简后的样本数据和粒子群算法优化模型系数，以期达到更高的精度和准确率。

1　数据来源

选取淄博矿井突水案例中 20 个典型底板突水资料作为原始标准样本数据[3]，其指标为含水性、水压、隔水层厚度、突水系数、导水性、构造发育、岩性组合、推进步距和工作面斜长 9 个因素。以这 20 个样本数据作为建模数据，具体数据见表 1。表 1 中的实际突水性一列里的数值“0”表示不突水，数值“1”表示突水。

2　粗糙集属性约简

粗糙集理论(Rough Sets，RS)是基于不可分辨关系的思想，具有很强的定性分析能力，能在不减少关键信息的前提下对数据进行约简，去除冗余属性和冗余样本，精简知识系统[5-6]。

表 1　淄博煤矿突水各影响因素与实际突水量(训练样本)

Table 1　Factors affecting water inrush in Zibo coal mine and actual yield of water inrush (training sample)

序号	含水性	水压 /MPa	隔水层厚度 /m	突水系数	导水性	构造发育	岩性组合	推进步距 /m	工作面斜长 /m	实际突水性	实际突水量 /($m^3 \cdot h^{-1}$)
1	0.5	1.48	33	0.64	0.5	0.8	0.5	30	142	0	0
2	0.5	0.85	26.6	0.51	0.5	0.5	0.5	30	80	0	0
3	0.5	3.32	63.32	0.59	0.5	0.5	0.5	30	150	0	0
4	0.5	2.22	35	0.74	0.5	0.5	0.5	30	60	0	0
5	0.5	3	40.32	0.99	0.5	0.5	0.5	30	130	0	0
6	0.5	1.08	27.79	0.54	0.5	0.5	0.5	30	85	0	0
7	0.5	1.46	29.7	0.82	0.5	0.3	0.5	30	120	0	0
8	0.5	1.34	30.43	0.7	0.5	0.3	0.5	30	120	0	0
9	0.5	1.14	25.85	0.71	0.5	0.3	0.5	30	100	0	0
10	0.5	1.52	25.5	0.72	0.5	0.5	0.5	30	50	0	0
11	0.5	1.52	28	0.84	0.5	1	0.5	5	142	1	156
12	0.5	0.87	17.66	1.14	0.5	0.5	0.5	30	100	1	120
13	0.5	1.13	24.1	0.81	0.5	0.5	0.5	32	120	1	54
14	0.5	1.91	25.5	1.03	0.8	0.5	0.5	40	85	1	330
15	0.5	2.19	28.4	0.98	0.5	0.5	0.5	45	66	1	15
16	0.5	1.86	26.4	0.88	0.5	0.5	0.5	30	45	1	51
17	0.5	1.48	29	0.87	0.5	0.3	0.5	30	110	1	12
18	0.5	2.18	24.5	1.28	0.5	0.5	0.5	30	80	1	60
19	0.5	4.25	49	1.15	0.5	0.5	0.5	84	120	1	195
20	0.5	2.86	48.02	0.79	0.5	0.5	0.5	30	75	1	60

注：含水性、突水系数、导水性、构造发育、岩性组合均为量化值，无量纲。

将表 1(决策表)中的数据作为信息系统，该系统可用四元组 $\boldsymbol{S}=(\boldsymbol{U}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{V}, \boldsymbol{F})$来表示，其中 $\boldsymbol{U}=\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{20}\}$为研究对象的集合，称为论域；$\boldsymbol{C}=$｛含水性、水压、隔水层厚度、突水系数、导水性、构造发育、岩性组合、推进步距、工作面斜长｝$=\{s_1, s_2, \cdots, s_9\}$为条件属性集；$\boldsymbol{D}=$｛无，有｝$=\{0, 1\}$为决策属性集(实际突水性)；$\boldsymbol{V}=\{\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots, \boldsymbol{V}_9\}$为属性的值域集，且 $\boldsymbol{V}_i$ 为属性 s_i 的值域；记 $\boldsymbol{R}=\boldsymbol{C}\cup\boldsymbol{D}$，则称 $\boldsymbol{R}$ 为属性集；$\boldsymbol{F}: \boldsymbol{U}\times\boldsymbol{R}\rightarrow\boldsymbol{V}$ 是一个信息函数，用 $\boldsymbol{F}$ 确定 $\boldsymbol{U}$ 中每个对象 $\boldsymbol{X}$ 的属性值。

对于 $\boldsymbol{R}$ 的任意属性子集 $\boldsymbol{B}$，其在 $\boldsymbol{U}$ 上的等价关系(等价集) $IND(\boldsymbol{B})=\{(x,y)\in U\times\boldsymbol{U} \mid \boldsymbol{F}(x,a)=\boldsymbol{F}(y,a), \forall a\in\boldsymbol{B}\}$。若 $(x,y)\in IND(\boldsymbol{B})$，则 x 与 y 相对 $\boldsymbol{B}$ 不可分辨。设 $\boldsymbol{Q}$ 是一个属性等价关系族(或称等价集族)，若 $IND(Q)=IND(Q-\{q\})$，则 q 为 Q 中可以被约去的知识。如果 $P=Q-\{q\}$ 是独立的，则 P 为 Q 的一个约简。Q 中所有不可约简的关系(知识)，即真正有用的部分，称为 Q 的知识“核”。粗糙集属性约简的具体步骤：

a. 由论域中的条件属性和决策属性构建二维决策表；

b. 采用粗糙集离散化算法将决策表中的条件属性和决策属性变量进行离散化(本文采用等频率

划分算法进行连续数据离散化，等频率区间数 $K=5$)。

c. 对离散化的新决策表按约简的决策规则进行属性简约。

以上过程由 MATLAB 软件完成[7]。对表 1 经过粗糙集属性约简后得到的约简属性为{突水系数、导水性、构造发育、岩性组合、推进步距、工作面斜长}＝{s_4，s_5，…，s_9}。以约简后的样本数据(新决策表)作为建模数据，见表 2。

表 2 粗糙集约简后的淄博煤矿突水各影响因素与实际突水量(训练样本)

Table 2 The influence factors of water inrush in Zibo coal mine after rough set reduction and the actual yield of water inrush (training sample)

序号	突水系数	导水性	构造发育	岩性组合	推进步距/m	工作面斜长/m	实际突水性	Logistic 模型识别结果
1	0.64	0.5	0.8	0.5	30	142	0	0
2	0.51	0.5	0.5	0.5	30	80	0	0
3	0.59	0.5	0.5	0.5	30	150	0	0
4	0.74	0.5	0.5	0.5	30	60	0	0
5	0.99	0.5	0.5	0.5	30	130	0	1
6	0.54	0.5	0.5	0.5	30	85	0	0
7	0.82	0.5	0.3	0.5	30	120	0	1
8	0.7	0.5	0.3	0.5	30	120	0	0
9	0.71	0.5	0.3	0.5	30	100	0	0
10	0.72	0.5	0.5	0.5	30	50	0	0
11	0.84	0.5	1	0.5	5	142	1	1
12	1.14	0.5	0.5	0.5	30	100	1	1
13	0.81	0.5	0.5	0.5	32	120	1	1
14	1.03	0.8	0.5	0.5	40	85	1	1
15	0.98	0.5	0.5	0.5	45	66	1	1
16	0.88	0.5	0.5	0.5	30	45	1	1
17	0.87	0.5	0.3	0.5	30	110	1	1
18	1.28	0.5	0.5	0.5	30	80	1	1
19	1.15	0.5	0.5	0.5	84	120	1	1
20	0.79	0.5	0.5	0.5	30	75	1	1

3 矿井突水的 Logistic 回归模型

设经粗糙集属性约简后的矿井突水 6 个诱发因素(突水系数、导水性、构造发育、岩性组合、推进步距、工作面斜长)的量化指标分别为 x_1，x_2，…，x_6(具体值见表 2)，p 为突水事件(模式)发生的概率，则 Logistic 变换 Logit(p)：

$$Y=\ln\frac{p}{1-p}\ (\text{或 logit}(p)=\ln\frac{p}{1-p}) \tag{1}$$

将概率区间[0，1]映射到区间$(-\infty,+\infty)$上，使得在任何自变量取值下，对 p 值的预测均有实际意义[8]。

大量实践证明，logit(p)往往和自变量呈线性关系。因此，以 logit(p)为因变量，建立包含 6 个自变量的 logistic 回归模型：

$$Y=\text{logit}(p)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_6x_6 \tag{2}$$

由式(2)可逆推得：

$$p=\frac{\exp(\beta_0+\beta_1x_1+\cdots+\beta_6x_6)}{1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_6x_6)} \tag{3}$$

式中 p 为概率或定性变量，或是具有二分性的变量。在本研究中，设矿井突水时的 $p_0=1$，未发生突水时的 $p_0=0$，但在实际预测分析时可约定：当 $p\in(0.5,1]$ 时可判定发生突水；当 $p\in[0,0.5)$ 时可判定未发生突水，即由式(3)算出的 p 值是标准值 p_0 的估计值(或近似值)。β_i 为 logistic 回归的偏回归系数，表示变量 x_i 对 Y 或 logit(p)的影响大小，β_0 为常数项。偏回归系数和常数项由粒子群算法确定。

3.1 基于粒子群算法的模型系数优化

模型系数 $\beta_0,\beta_1,\cdots,\beta_6$ 对模型精度有较大影响，本文采用粒子群算法(PSO)优化模型系数。

粒子群优化算法(Particle Swarm Optimization，PSO)是 Kennedy 和 Eberhart 受鸟群觅食行为启发于 1995 年提出的一种全局优化算法[9-10]。PSO 算法(一种仿生算法)具有搜索能力强、收敛速度快、设置参数少、程序易实现和无需梯度信息等特点。所有的粒子(搜索空间中的鸟，本文中粒子为待优化参数序列)都有一个由被优化函数决定的适应值和一个决

定它们运动方向和运动距离的速度，问题的解就是搜索空间中的一只鸟的位置。PSO 算法的主要特点为：**a.** 每一粒子都被赋予了初始随机速度并在解空间中流动；**b.** 个体具有记忆功能；**c.** 个体的进化主要是它本身的飞行经验以及同伴的飞行经验进行动态调整，通过迭代找到最优解。在每一次迭代过程中，粒子通过追逐两个极值来更新自己的位置。一个是粒子自身所找到的当前最优解，这个解称为个体极值 P_{best}；另一个是整个群体当前找到的最优解，这个解称为全局极值 P_{gbest}。为了提高 PSO 算法的搜索能力，防止早熟或避免陷入局部最优现象，借鉴遗传算法中的变异思想，在普通粒子群算法的基础上引入了简单变异算子，使 PSO 算法能够跳出局部极小点。模型系数优化算法主要步骤如下[11]：

a. 随机初始化粒子群的位置和速度，设定学习因子 c_1、c_2，最大迭代次数 M，种群规模 N 以及待优化系数的取值范围。

b. 利用每个微粒对应的向量(β_0，β_1，…，β_6)分别建立 logistic 回归模型模型，计算每个微粒的适应度值，将当前每个微粒的位置和适应值存储在各微粒对应的 P_i 中，将所有 P_i 中适应值最优的个体的位置和适应值存储在 P_g 中。适应值 f 计算式为：

$$f=\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i-p_i)^2 \tag{4}$$

式中 n 为训练样本数($n=20$)，$\hat{p}_i$ 和 $p_i\in\{0,1\}$ 分别为第 i 个训练样本突水概率的计算值和实际值(标准值)。

c. 更新各微粒的速度和位置，

$$\begin{cases} v_{i,j}(t+1)=w\cdot v_{i,j}(t)+c_1\text{rand}()[p_{i,j}-x_{i,j}(t)]+ \\ \qquad\qquad c_2\text{rand}()[p_{g,j}-x_{i,j}(t)] \\ x_{i,j}(t+1)=x_{i,j}(t)+v_{i,j}(t+1) \end{cases} \tag{5}$$

式中 $v_{i,j}(t+1)$ 表示第 i 个粒子在 $t+1$ 次迭代中第 $j(j=1,2,\cdots,7)$ 维上的速度；w 为惯性权重，rand() 为 0～1 之间的随机数。此外，为使粒子速度不致过大，可设置速度上下限，即 $v_{i,j}\in[-v_{\max},v_{\max}]$，$v_{\max}$ 是之前设定的最大速率(边界值)，t 为当前迭代次数。

d. 对每个微粒的适应值及其经历过的最好位置进行比较，如果较好将其作为当前最好位置。比较当前所有 P_i 和 P_g 值，更新 P_g。

e. 若满足停止条件(预设的运算精度或迭代次数)，搜索停止输出结果，否则，返回 **c** 继续搜索。

f. 利用最优参数 $\beta_0,\beta_1,\cdots,\beta_6$ 建立模型并进行突水模式诊断。

3.2　算例分析

首先设置粒子群算法中的相关参数：$c_1=2$，$c_2=2$，$v_{\max}=1$，种群规模为 $N=50$，最大迭代次数 $M=100$。当前惯性权重为

$$w=w_{\max}-\text{iter}\times\frac{w_{\max}-w_{\min}}{\text{iter}_{\max}} \tag{6}$$

式中 $w_{\max}=0.9$，$w_{\min}=0.4$；iter 为当前迭代次数，$\text{iter}_{\max}=M=100$。待估参数 $\beta_0,\beta_1,\cdots,\beta_6$ 的取值范围为 $[-10,10]$。利用表 2 中的建模样本数据及式(4)设计适应度函数 f 的计算式(由 MATLAB 编程实现，在此略去)。运算执行后，辨识误差函数 f (个体适应度函数)的优化过程如图 1 所示。从图 1 可看出只需经过 50 次的迭代函数便收敛于 0.3，说明收敛速度非常快。

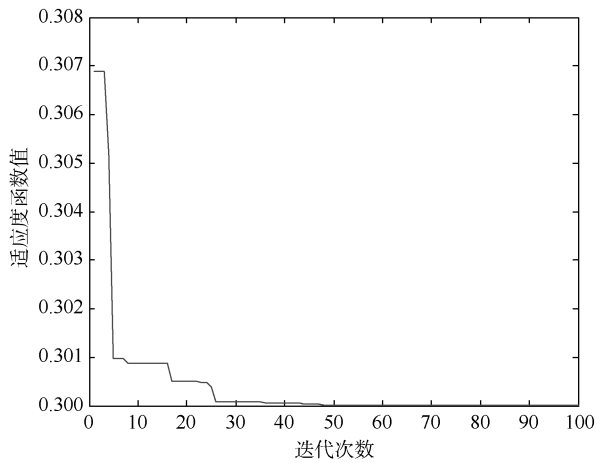

图 1　适应度函数变化曲线

Fig.1　Fitness function curve

完成辨识后输出的最优模型系数向量为 $[\beta_0,\beta_1,\cdots,\beta_6]$=[–10.962 5, 13.965 2, 0.972 1, 1.618 3, –2.588 9, –0.002 3, 0.000 0]，对应的 Logistic 回归模型如下：

$$p=\frac{\exp(-10.9625+13.9652x_1+0.9721x_2+1.6183x_3-2.5889x_4-0.0023x_5+0x_6)}{1+\exp(-10.9625+13.9652x_1+0.9721x_2+1.6183x_3-2.5889x_4-0.0023x_5+0x_6)} \tag{7}$$

利用式(7)对训练样本反向检验，计算出的 p 值向量为(0.167 7, 0.019 7, 0.058 1, 0.333 2, 0.942 7, 0.029 7, 0.525 4, 0.171 6, 0.192 3, 0.274 2, 0.828 0, 0.992 6, 0.569 9, 0.974 1, 0.932 4, 0.779 2, 0.689 9, 0.998 9, 0.992 7, 0.501 2)，根据 $p>0.5$ 判定为突水，$p<0.5$ 判定为无突水的规则得出训练样本的突水模式，结果见表 2，正确率为 90%。对测试样本[3](样本属性与表 2 保持一致)的诊断结果见表 3。

将表 3 中的每个测试样本 6 个属性值依次代入式(7)便可得到 Logistic 回归模型 p 的预测值向量(0.843 2, 0.158 3, 0.991 5, 1.000 0, 1.000 0, 0.970 5)，根据 $p>0.5$ 判定为突水，$p<0.5$ 判定为无突水的规则，得到测

表 3 矿井突水测试样本的诊断结果及比较

Table 3 Diagnostic results and comparison of test samples of water inrush

测试样本序列	矿井突水特征						属性约简的 Logistic 回归模型诊断结果	非属性约简的 Logistic 回归模型诊断结果	实际结果
	x_1	x_2	x_3	x_4	x_5	x_6			
1	0.91	0.5	0.5	0.5	30	120	1	0	1
2	0.67	0.5	0.5	0.5	30	75	0	0	0
3	1.13	0.5	0.5	0.5	30	110	1	1	1
4	4	0.8	0.8	0.8	28	70	1	1	1
5	2.5	0.5	0.5	0.5	12	110	1	1	1
6	1.04	0.5	0.5	0.5	30	28	1	1	1

试样本的诊断结果见表 3 第 8 列，正确率为 100%。

使用全属性(9 个)建模样本集(表 1)建立 Logistic 回归模型(仍采用粒子群算法优化模型系数，粒子群算法的参数设置不变)，并对建模样本集反向检验，求出的 p 值向量(0.001 2, 0.011 7, 0.000 0, 0.001 4, 0.006 3, 0.011 5, 0.277 3, 0.053 9, 0.495 6, 0.293 4, 0.001 5, 0.999 9, 0.909 0, 0.995 7, 0.980 0, 0.605 9, 0.455 4, 0.999 2, 0.915 8, 0.000 0)。根据 p 值大小得出突水判定结果为(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0)，正确率为 85%；求出的全属性测试样本(另 3 个属性值见文献[3])的 p 值向量为(0.000 0, 0.000 0, 0.988 4, 1.000 0, 1.000 0, 0.952 9)，诊断结果为(0, 0, 1, 1, 1, 1)(表 3 中的第 9 列)，正确率为 66.7%。可见，基于粗糙集的 Logistic 回归模型具有很高的精确度，好于属性非约简的 Logistic 回归模型。

4 结 语

矿井突水模式识别是一个非正态、非线性和高维数据处理问题，也是二分类问题。利用粗糙集理论对数据实施降维，并建立二项分类 Logistic 回归突水模式识别模型，该模型将影响因素的线性关系与因变量(二分类变量)之间的非线性关系式有机地结合，克服了线性回归分析解决二分类问题的不足。利用粒子群算法优化模型参数，通过对建模样本和测试样本的突水模式检验和诊断结果来看，基于粗糙集的 Logistic 回归模型具有很高的精确度，效果好于属性非约简的 Logistic 回归模型，为矿井突水模式识别提供了一种新思路、新方法。

参考文献

[1] 张立新，李长洪，赵宇. 矿井突水预测研究现状及发展趋势[J]. 中国矿业，2009，18(1)：88–90.

[2] 张自政，杨勇，田立娇，等. 模糊评价分类模型在矿井底板突水判别中的应用[J]. 矿业安全与环保，2010，37(6)：41–43.

[3] 张文泉. 矿井底板突水灾害的动态机理及综合判测和预报软件开发研究[D]. 青岛：山东科技大学，2004：164–166.

[4] 王济川，郭志刚. Logistic 回归模型—方法与应用[M]. 北京：高等教育出版社，2001.

[5] 张小红，裴道武，代建华，等. 模糊数学与 Rough 集理论[M]. 北京：清华大学出版社，2013：259–274.

[6] 赵军，张显跃. 基于粗集理论的数据离散化技术研究[J]. 重庆邮电学院学报，2006，18(6)：752–757.

[7] 许国根，贾瑛. 模式识别与智能计算的 MATLAB 实现[M]. 北京：北京航空航天大学出版社，2012：149–160.

[8] 张文彤，钟去飞. IBM SPSS 数据分析与挖掘实战案例精粹[M]. 北京：清华大学出版社，2013：168–169.

[9] EBERHART R C．KENNEDY J．A new optimizer using particle swarm theory[C]//in：Proc．the sixth international symposium on Micro Machine and Human Science．Nagoya，Japan：[s.n.]，1995：39–43.

[10] [KENNEDY J，EBERHART R C，SHI Y．Swarm intelligence[M]. San Francisco：Morgan Kaufmann Publishers，2001.

[11] 王江荣. 基于粒子群算法的自回归加权马尔可夫链的负荷预测[J]. 工业仪表与自动化装置，2014(1)：113–117.

(责任编辑 张宏)