一种处理丢失数据的混合算法

任仲晟
（福建师范大学数学与计算机科学学院 福建福州 350007）

摘 要：数据挖掘经常会碰到缺失数据。本文结合回归和决策树的优点提出一种混合算法能较好地处理文本和连续性数据，同时考虑到当文本型数据类别很大时，决策树处理方法效果不佳，提出了一种C4.5的改进算法。

关键字：数据丢失 回归 聚类 C4.5
中图分类号：TP301 文献标识码：A 文章编号：1007-0916(2010)09-0024-02

A Hybrid Algorithm dealing with Missing Data

Abstract: In this paper, by combining advantages of regression and decision tree, we provide a hybrid algorithm which can process data of text type and continuous data. Also an improved algorithm based on C4.5 is offered.

Key Word: missing data, regression, clustering, C4.5

1 引言

根据文献[1]的研究，在数据挖掘过程中20%的时间用于目标识别，60%的时间用于数据准备，数据挖掘和知识分析都为10%，人们要将超过50%的精力放在数据预处理上。

数据丢失是数据预处理中面临的一大问题。已有文献对缺失数据的处理提出了多种方法。例如，可以直接忽略含有缺失属性值的记录，利用均值或同类别均值填补缺失值，利用回归分析、贝叶斯计算公式或决策树推断出该条记录特定属性最大可能的取值用以填补缺失值等等。本文在已有研究基础上，提出了一种新的混合算法。

2 常用丢失数据处理算法

2.1 mean-mode算法

该方法的思路是使用全部数据集重，数值属性的均值，和符号属性中出现最频繁的值来填充缺失数据。具体算法如下：(1)对于具有缺失数据的每个属性。(2)如果该属性是数值得，计算所有完整数据的均值。(3)否则，如果该属性是符号，统计每个取值出现的次数，取出频率最高的那个属性值。(4)将计算得到的值填充缺失数据。

2.2 基于聚类的处理算法

如果说mean-mode算法是从纵向即按照列的角度处理丢失数据，那么聚类算法则是按照行向的角度处理丢失数据，把这二者算法结合起来就是从行和列的角度来考虑丢失数据。具体算法如下：(1)用mean-mode填充丢失的数据。(2)利用k-means 聚类算法把所有的样本中数值型数据聚类。(3)利用k-mode聚类算法把所有的样本中文本型数据聚类。(4)利用xie-beni验证所得的最优聚类数的聚类中心填充到丢失的数据中。

2.3 基于聚类的回归模型处理算法

为了提高回归模型预测的效率，可以将具有相似特征的样本聚类(样本分层)，然后对每一类分别建立回归模型，并进行预测和替代缺失部分数据。聚类分析的方法有多种，本文采用基于k-means算法的快速聚类算法。具体算法如下：(1)给定数据集中的缺失值，基于均值替代生成完整的数据集。(2)利用填补完的数据聚类，对每一类建立回归方程。(3)根据上一步的回归分析，计算出缺失数据的值，并填补缺失数据。

表1

处理方法	0-0.05	0.05-1	0.1-0.2	0.2-0.3	>0.3	总计
Mean_mode	13.16%	18.42%	18.42%	17.54%	32.46%	0.26
cluster	14.29%	32.14%	7.14%	3.57%	42.86%	0.24
聚类回归	26.92%	15.38%	30.77%	15.38%	11.54%	0.12

表2

	聚类	C4.5 决策树算法	C4.5 改进算法
5%丢失	0.24	0.18	0.14
10%丢失	0.29	0.20	0.18
20%丢失	0.38	0.36	0.32

表3

	聚类回归		C4.5 决策树算法		决策-回归	
	数值	文本	数值	文本	数值	文本
5%丢失	0.21	0.34	0.33	0.2	0.16	0.2
10%丢失	0.25	0.43	0.37	0.24	0.22	0.24
20%丢失	0.33	0.52	0.43	0.32	0.29	0.32

3 聚类C4.5算法

在处理数据丢失时，决策树在处理文本型数据方面优于回归，但当文本型数据类别较多时则错误率增大。我们提出一种改进的C4.5算法(聚类C4.5算法)。算法思想如下：

我们首先对要预测数据所在属性按属性值相等简单地分类(A1，A2…Ak)，对每个Ai进行聚类得到一个聚类中心，再对这些聚类中心进行聚类(5至10类)，由聚类得出属于同一类的所有属性值，把它当成纯节点的判定条件来建决策树，得到树T1，根据树T1判定丢失数据所属类别(它有多个属性值a1，a2…ai)，再对属性值属于a1，a2…ai的所有数据进行建决策树T2，再根据树T2判定丢失数据所属类别，并填入相应数据。

4 回归和决策树混合算法

在建回归方程时，若元组个数很大时填补的数据效果不好。为此，我们提出决策树和回归混合算法。具体算法如下：(1)if 丢失的数据是文本型。根据丢失数据所在属性建决策树T1，根据T1找出该数据所属的叶节点，并填充。(2)分别对每个文本属性作为最终分类属性建决策树T1，T2……Ti。(3)对训练数据集的连续型数据做随机丢失作为新的训练数据集。(4)比较基于每个Ti分类后得出的回归方程对训练数据集中连续型数据准确率。(5)选择4中得出的具有最优准确率的树Ti作为最终的分类树建回归方程。(6)通过回归方程计算出丢失

(下转26页)

色(决定做什么)分开。

视窗:由JSP建立,struts包含扩展自定义标签库,可以简化创建完全国际化用户界面的过程。

控制器:struts中,基本的控制器组件是ActionServlet类中的实例servelt,实际使用的servlet在配置文件中由一组映射(由ActionMapping类进行描述)进行定义。

2.2 系统用例图

从考生角度的系统用例图如图2所示:

3 性能测试

使用HP公司的LoadRunner进行性能测试,首先要通过Virtual User Generator来完成测试脚本录制与开发工作,再进行场景创建与执行,最后进行性能测试结果分析。

3.1 脚本录制

使用LoadRunner的脚本创建与录制功能进行考生注册、考生身份验证、考试科目选择、进行考试和成绩查询等操作的脚本录制工作。其中考生登录验证过程的脚本录制如图3所示:

3.2 场景创建与执行

创建不同数量的虚拟考生进行登录验证操作的并发执行的测试结果如表1所示:

3.3 测试结果分析

经过测试发现最初采用的jdbc-odbc桥连接的方式连接数据库随着并发用户增多SQL语句执行效率迅速降低,为解决这一问题采用连接池技术进行数据库的连接。为数据库连接建立一个"缓冲池"。预先在缓冲池中放入一定数量的连接,当需要建立数据库连接时,只需从"缓冲池"中取出一个,使用完毕之后再放回去。通过设定连接池最大连接数来防止系统无尽的与数据库连接,更为重要的是可以通过连接池的管理机制监视数据库的连接的数量、使用情况,为系统开发、测试及性能调整提供依据。改进后的程序测试结果如表2所示:

4 结语

虽然经过功能测试和集成测试发现客户端软件与服务器端软件功能上均符合设计要求且各模块之间通信状态与预期相符,也未必能够满足性能要求,虽然从单元测试起,每一测试步骤都包含性能测试,但只有当系统真正集成之后,在真实环境中才能全面、可靠地测试出运行时的性能。进行性能测试是发现系统瓶颈、提高软件质量的有效途径,在软件开发的整个生命周期中占有十分重要的地位。

表1 考生进行登录验证操作测试结果

并发用户数	响应延时	不能成功登录用户数	服务器CPU利用率	服务器内存利用率
1	0.01s	0	20%	25%
50	0.1s	0	22%	25%
100	1.2s	3	40%	55%
500	3.8s	35	60%	65%
1000	6.5s	140	90%	85%

表2 改进后的考生进行登录验证操作测试结果

并发用户数	响应延时	不能成功登录用户数	服务器CPU利用率	服务器内存利用率
1	0.01s	0	20%	25%
50	0.05s	0	22%	25%
100	0.2s	0	35%	45%
500	0.5s	0	55%	55%
1000	0.8s	8	70%	65%

参考文献

[1]Booch Grady,Rumbaugh Jame, Jacbson Ivar.UML用户指南[M].邵维忠.麻志毅.张文娟,等译.北京:机械工业出版社.2001.

[2]程绍英,刘建华,金成姬.LoadRunner性能测试实战[M].电子工业出版社2007.

[3]Paul C.Jorgensen.Software Testing[M].韩柯,杜旭涛,译.机械工业出版社.2003.

[4]郑人杰,马素霞,麻志毅.软件工程[M].人民邮电出版社.2009.

基金项目:黑龙江省教育厅科学技术研究项目(11533050)。

(上接24页)

数据。

5 实验结果

5.1 数值型数据

0-0.05:表示丢失数据与原始数据的差值在0-0.05范围内丢失数据个数所占总丢失数据个数的百分比,余同(表1)。

5.2 文本数据

文本型数据丢失的误差率(表2)。

5.3 混合型数据

混合型数据丢失的误差率(表3)。

6 结语

本文采用mean-mode,聚类,聚类回归,C4.5决策树,改进的C4.5算法,决策树+回归来衡量填补的丢失数据值效果,通过实验数据分析可以看出(1)对处理连续型数据丢失,聚类回归效果最好,原因是利用聚类和线性回归把各个属性间关系关联起来,从而能更好预测丢失值。(2)对处理文本型数据丢失,C4.5决策树算法效果比较好,因为C4.5考虑到属性间的联系,而聚类没有考虑,而改进的C4.5主要对文本型数据分类比较多效果较优。(3)对处理混合型数据,决策+回归效果较好,因为决策树善于处理离散型变量,而回归善于处理连续型变量。

参考文献

[1]唐华松,姚耀文.数据挖掘中决策树算法的探讨.计算机应用研究.

[2]谭旭,王丽珍,卓明.利用决策树发掘分类规则的算法研究.云南大学学报(自然科学版).

[3]谢志鹏,张卿.基于粗糙集合理论的决策树生成.计算机工程与应用.

[4]Jin yong-Jin, Imputation for Missing Values,Mathematic Statistic and Manage.

一种处理丢失数据的混合算法

作者： 任仲晟
作者单位： 福建师范大学数学与计算机科学学院,福建福州,350007
刊名： 数字技术与应用
英文刊名： DIGITAL TECHNOLOGY AND APPLICATION
年，卷(期)： 2010，(9)
被引用次数： 0次

参考文献(4条)

1. 唐华松,姚耀文.效据挖掘中决策树算法的探讨.计算机应用研究.

2. 谭旭,王丽珍,卓明.利用决策树发掘分类规则的算法研究.云南大学学报(自然科学版).

3. 谢志鹏,张卿.基于粗糙集合理论的决策树生成.计算机工程与应用.

4. Jin yong-Jin,Imputation for Missing Values,Mathematic Statistic and Manage.

相似文献(9条)

1. 期刊论文 邸亚洲.秦永元.尚希良.曲建岭.DI Ya-zhou.QIN Yong-yuan.SHANG Xi-liang.Qu Jian-ling 基于多项式回归算法的飞参记录数据预处理研究 -测控技术2008,27(4)

提出利用多项式回归算法对飞参记录数据存在的随机量测误差、野点以及数据丢失等现象进行有效的数据预处理,算法在消除量测误差、剔除和补正野点、补充丢失的数据及数据平滑等方面均具有较高的精度和可靠性并已有效地应用在多型飞机飞参记录数据预处理工作中.利用该算法在Matlab环境下对飞参记录的航姿系统俯仰通道部分数据预处理过程进行了仿真.

2. 学位论文 李荣 基于删失试验的贝叶斯生存回归模型及其应用 2006

随着科学技术的发展，产品的生存状况愈来愈受到人们的重视。由于产品的寿命是一个随机现象，所以确定一种产品的可靠性指标最后都归结为一个统计推断问题，为了弄清被测产品的寿命分布，求出各项可靠性指标，研究产品的失效机理以便对提高产品可靠性提出建议，常常需要进行寿命试验。

在用这些统计方法处理实际问题时，常会遇到数据删失问题，譬如在产品寿命试验中，由于试验设备，观测手段或有其他方面的困难造成某些试验数据丢失或未观测到的现象等，这样我们得到的是删失数据。如何对删失试验产生的数据进行统计分析，在生存分析中是一个非常重要的领域。

在删失试验中，数据丢失场合下，若仅根据所剩数据提供的信息，人们对产品的各项指标的统计推断结果的可信性会受到影响。本文讨论了生存分析中的三个模型：极值回归模型、威布尔回归模型和分段指数共享异质模型。在充分利用已有信息的情况下，给出参数的贝叶斯估计。但计算很是困难，原因就是删失数据后样本的似然函数的形式很复杂。为了计算的简便，利用Gibbs抽样得出参数的后验分布。利用WinBUGS软件进行数据仿真分析。对三个模型分别设计出WinBUGS程序和模型有非循环有向分析图，得出后验参数相关统计估计量和相关的统计诊断图。说明三个模型在生存分析中的有效性和可靠性。

3. 期刊论文 田晓红.林友明.TIAN Xiao-hong.LIN You-ming Landsat-7 缝隙数据恢复的算法研究 -计算机仿真2007,24(12)

扫描行校正器异常造成了Landsat-7图像数据丢失(称之为缝隙数据),由于数据本身仍保持了良好的辐射和几何特性,具有较好的可用性,因此研究缝隙数据的恢复算法将具有较好的理论价值和应用前景.目的就是通过仿真实验寻找一种较好的恢复算法.首先介绍了课题背景以及现有的算法—全局直方图匹配法和局部直方图匹配法,并在局部直方图匹配算法的基础上提出了自适应局部回归匹配算法.最后对同一幅缝隙图像用这三种算法来实现恢复,通过比较均方根误差和运行速度,论证了自适应局部回归匹配算法在精度方面要优于全局直方图匹配算法和局部直方图匹配算法,算法复杂度要优于局部直方图匹配算法,具有很高的可行性.

4. 学位论文 高红 基于结构张量的核回归非均匀插值算法及其在图像处理中的应用 2009

近年来，随着信息技术的发展，数字图像处理的应用越来越广泛，而插值作为图像处理的一个重要手段，有着很重要的作用。在图像处理中很多问题都可以转化为插值问题来实现，如图像放大、图像去噪、超分辨率重建等问题。传统的插值方法一般是针对均匀数据而言的，但在实际操作中，由于各方面的误差，如运动模糊，数据丢失等原因，样本点往往是不规则的，因此对非均匀插值算法的研究具有很重要的实际意义。

本文首先重点介绍了核回归的非均匀插值算法，针对核回归中核函数选择的缺陷，提出了一种基于结构张量的核回归非均匀插值算法。同时研究了改进的算法在图像放大、图像去噪、超分辨率重建等方面的应用。论文主要的研究成果如下：

(1)提出基于结构张量的核回归非均匀插值算法。首先介绍了最新的用核函数回归的方法进行非均匀插值的算法，并分析了鲁棒、自适应的Bilateral KernelRegression和Steering Kernel Regression两种核函数回归的缺点和局限性。然后提出了一种基于结构张量的核函数回归算法，用结构张量来估计图像的结构信息和灰度信息，使得核函数在边缘区域能够自适应，从而使得非均匀插值得到很好的效果。最后用改进算法进行图像的随机采样的插值恢复，实验表明基于结构张量的核回归算法比较简单，减少了计算量，并且效果很好。

(2)改进的非均匀插值算法在图像处理中的应用。首先对图像放大、图像去噪、超分辨率重建等问题进行描述，并将这些问题转化为插值问题，建立恢复模型。然后用改进的算法来实现图像的放大、去噪与超分辨率重建。最后通过实验说明改进算法在估计图像的结构与纹理上具有很好的效果。

5. 学位论文 田晓红 Landsat-7 ETM+SLC-OFF图像缝隙数据修复算法研究 2007

美国陆地卫星7号(Landsat-7)于1999年4月15日由美国航天局(NASA)发射升空，其携带的主要传感器为增强型主题成像仪(ETM+)。2003年5月31日，Landsat-7 ETM+机载扫描行校正器(Scan Lines Corrector，简称SLC)突然发生故障，导致获取的图像出现数据重叠和大约25%的数据丢失，但是数据本身仍然保持了良好的几何特性和辐射特性。SLC异常的确大大影响和限制了数据的使用，但是数据仍然可以成功的运用到很多科学领域。论文正是基于这样的背景，提出了一些修复异常图像数据的算法，修复过的图像数据在某些领域的可用性会大大提高。

论文首先介绍了Landsat-7扫描行校正器的工作模式以及缝隙产生的原因，已有的基于重采样的图像数据填充算法无法恢复出缝隙内所有的像素值，而且恢复部分也是以牺牲了图像分辨率为代价的。在此基础上论文提出了用一景或多景SLC-ON图像(SLC异常之前完整图像)去填充SLC-OFF缝隙图像(SLC异常之后图像)的方法，采用了全局直方图匹配算法、局部直方图匹配算法和自适应局部回归算法。实验证明局部直方图匹配算法和自适应局部回归算法都取得了很好的效果。论文还介绍了用相邻周期的多景SLC-OFF图像相互填充融合的算法，这个阶段采用的是自适应局部回归算法，实验证明，因为数据接收时间上更接近，所以恢复出来的图像效果比用SLC-ON图像填充SLC-OFF图像的效果要好。

为了能充分利用SLC-OFF缝隙图像自身的光谱像素信息，论文又提出了基于图像分割和纹理合成的恢复算法。先用分形网络演化算法(FNEA)对SLC-ON图

像进行分割，将分割后的模型覆盖到SLC-OFF图像中，以便确定缝隙中各个像素分别属于哪一块。然后在每个块中用徐晓刚多种子纹理合成算法对缝隙部分进行填充。实验表明，这种方法因充分利用了SLC-OFF图像本身的光谱像素信息使得填充精度很高，大大提高了SLC-OFF图像数据的可用性。

6.期刊论文 卢新海.边馥苓.LU Xinhai.BIAN Fuling 大冶铁矿露天采场位移监测数据预处理技术 -地理空间信息 2008,6(6)

作为预警与决策主要依据的位移监测数据,存在非均匀时段观测、数据丢失和因观测误差剔除等问题.采用一元线型回归、三次样条的方法对观测数据进行预处理,比较处理结论,为预测判据提供科学的基础数据.

7.学位论文 李俊 基于HVS和CIRC错误校正的自适应图像水印研究 2008

随着计算机技术、数字多媒体技术和Intemet技术的飞速发展，各种形式的多媒体数字作品极易被非法复制、篡改和传播，因而如何对数字作品进行保护已成为一个相当重要而又富有挑战性的研究课题。目前，数字水印技术是保护多媒体数字作品版权的一种非常有效的方法。对数字水印技术的研究有着重要的理论和现实意义。

本文主要研究了图像数字水印技术的算法和应用，以静态图像为研究对象，其主要研究成果及核心部分归纳如下：

1.本文提出一种基于支持向量回归机和人类视觉系统的图像水印算法。由于水印的嵌入强度越大，水印的鲁棒性就越好，所以为了保证水印不可见的同时提高每个水印的嵌入强度，本文通过考虑人类视觉特性，根据图像的局部特征选择相适应的水印嵌入强度。人类视觉特性是很难解析和描述的，支持向量回归机具有非常好的非线性拟合能力，能够成功的应用于函数逼近方面，所以本文利用支持向量回归机模拟人类视觉特性，根据图像的局部特征推测出每个水印的最适应的嵌入强度，使每个水印嵌入强度在确保数字水印不可见的同时最大。为了提高水印的鲁棒性，本文又根据图像的离散小波分解原理，借助嵌入零树小波编码思想，构造出图像经过小波分解后的重要小波系数，然后在重要的小波系数上嵌入水印。

2.含有水印的数字图像对于常见的数据操作或攻击会导致数据丢失，因此提取出的水印不可能完全正确。本文通过对现有的有关信息错误检测和校正算法的研究，提出了一种基于交叉交插里德一索洛蒙码(CIRC)的水印信息错误检测和校正算法。该算法根据CRC检测原理对提取出的水印信息进行检测是否错误，如果提取出的水印信息错误则采用交叉交插里德一索洛蒙码对错误的数据进行纠错。实验结果表明该算法可以有效地提高水印的鲁棒性。

8.期刊论文 肖英.李明亮.XIAO Ying.LI Mingliang 缺少观测数据下信号谱分析 -现代电子技术2007,30(8)

提出了一种处理时间序列中出现数据丢失时的信号谱估计的方法.这时观测所得到的,不再是连续等间隔的时间序列,而是多个数据段,所要进行的即是对这些分段数据的自回归模型的估计.该方法基于标准Burg谱估计算法提出,算法可以建立一个同时适用于各个分段数据的统一的信号模型.在仿真部分的结果显示,与直接使用均值方法进行谱估计相比较,分段Burg算法偏差更小,谱估计更精确.

9.学位论文 王文丰 分布式存储系统中高性能和高可靠性问题的研究 2009

计算机技术和宽带网络技术的迅猛发展以及存储市场的巨大需求，极大地推动了分布式存储技术的进步，同时也给现有的存储系统不断地提出各种新的要求。对分布式存储系统而言，系统应该能够提供始终如一的、高质量的存储服务，尽量降低由于网络环境的动态性和不可预知性以及热点数据访问等原因对系统服务的可靠性和服务性能所造成的影响。另外，数据的重要性也决定了高可靠性是分布式存储系统的基本目标之一。在信息化程度越来越高的今天，数据丢失已经变得不可忍受，因为一些重要信息的丢失往往会给企业带来巨大的经济损失。因此，如何实现高性能、高可靠的存储服务是当今存储系统中亟待解决的关键问题。

然而，纵观现有的各种分布式存储系统，发现它们在服务性能和可靠性方面仍然存在一些问题，主要表现在：1）目前大多数存储系统普遍存在主服务器性能瓶颈和单点失效问题，由此容易造成系统服务的不可靠和服务性能低下。虽然心跳机制能够在一定程度上降低单点失效发生的可能性，但在系统服务的可用性和服务性能等方面仍然存在着不足之处。2）现有的任务调度算法往往追求单一的调度目标，虽然可以使得加权总完成时间最优，但是在任务的平均周转时间方面考虑不足，并且在某些情况下可能会导致“饥饿”现象。3）现有的大多数副本管理策略主要根据节点对文件的访问频率或者系统总的请求响应时间来选择合适的副本放置节点，缺乏对单个请求的响应时间要求进行考虑，这可能会造成部分用户的请求响应时间过长。4）目前关于存储系统可靠性方面的研究主要是围绕数据冗余方法的研究而展开，而较少关注甚至忽略了存储资源分配方案对系统整体可靠性的影响。现有的存储资源分配方案虽然简单直观，但在文件大小和文件的重要性对系统整体可靠性的影响方面考虑不足。

本文针对现有的存储系统在服务性能和可靠性方面所存在的一些问题，分别从系统服务模型、任务调度算法、副本分发机制以及存储资源分配方案四个方面进行了系统而深入的研究，取得了若干创新性成果。

本文的主要研究工作和创新性成果体现在以下几个方面：

1.针对现有存储系统在服务的可靠性和服务性能等方面存在的不足，首先引入一种动态k叉树结构，给出了动态k叉树的相关定义和算法。然后在此基础上，提出了一种基于系统负载的轮流服务模型--ASSL（Alternate Service based on System Load）。在ASSL模型中，首先采用自回归负载预测模型来预测节点的负载以及过载发生的时间，这样，可以在节点过载发生之前采取主动防范措施（选举新的服务节点），从而提高了系统服务的可用性和服务性能；其次，通过采用基于选举域划分的多机心跳机制方法，减少了节点失效的检测时间和主节点的通信量；最后，为了进一步降低选举开销，对主节点过载和失效两种情形分别采取不同的选举机制。理论分析和实验结果表明，该模型对提高系统服务的可靠性、可用性以及服务性能是有效的。

2.分析了现有任务调度算法MTWCT（Minimize Total Weighted Completion Time）存在的不足，在此基础上提出了一种改进的优化任务调度算法E-TWCT（Enhanced TWCT），并给出了ρ因子的划分规则、△ρ和△t的临界值的设定方法以及E-TWCT算法的调度策略，同时进行了算法复杂度分析。实验结果表明，本文提出的算法不仅能够有效地消除“饥饿”现象，而且能够获得和MTWCT算法相同或者更优的平均周转时间，并且加权总完成时间和MTWCT算法相当。

3.分析了现有副本管理策略存在的不足，根据副本分发方案所需满足的目标要求，建立了一种基于响应时间度量的动态副本分发模型，并设计了求解该模型的遗传算法。实例分析表明，本文求解的副本分发方案RPRTM（Replica Placement based on Response Time Measure）能够在满足各个节点的单个请求的响应时间要求的同时使得系统所需创建的副本数最小化，而且在最大程度上缩短了系统总的请求响应时间，提高了系统整体服务性能。

4.针对现有存储资源分配方案存在的不足，提出了文件优先级比重的概念，充分考虑了文件大小和文件重要性对系统整体可靠性收益的影响。在此基础上，研究了有限资源条件下如何对多个大小不同、重要性不同的文件进行资源分配的问题。建立了一种非线性整数规划模型，求解并得出了能够使得系统整体可靠性收益达到最大的理论最优资源分配方案和可行最优资源分配方案，同时给出了相关的理论推导和证明。实验结果表明，相比现有的资源分配方案而言，本文求解的资源分配方案能够在相同存储资源条件下获得更高的系统可靠性收益。