第七届数学中国数学建模网络挑战赛
地址：数学中国数学建模网络挑战赛组委会　　网址：www.tzmcm.cn
电话：0471-4969085　　邮编：010021　　Email：2014@tzmcm.cn

第七届“认证杯”数学中国

数学建模网络挑战赛
承　诺　书

我们仔细阅读了第七届“认证杯”数学中国数学建模网络挑战赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们接受相应处理结果。

我们允许数学中国网站(www.madio.net)公布论文，以供网友之间学习交流，数学中国网站以非商业目的的论文交流不需要提前取得我们的同意。

我们的参赛队号为：2351

参赛队员　(签名)　：

队员 1：呼和

队员 2：张雅丽

队员 3：王政

参赛队教练员 (签名)：　马壮

参赛队伍组别：本科

第七届“认证杯”数学中国

第七届数学中国数学建模网络挑战赛

地址：数学中国数学建模网络挑战赛组委会 网址：www.tzmcm.cn
电话：0471-4969085 邮编：010021 Email：2014@tzmcm.cn

数学建模网络挑战赛
编 号 专 用 页

参赛队伍的参赛队号：（请各个参赛队提前填写好）：

2351

竞赛统一编号（由竞赛组委会送至评委团前编号）：

竞赛评阅编号（由竞赛评委团评阅前进行编号）：

第七届数学中国数学建模网络挑战赛
地址：数学中国数学建模网络挑战赛组委会　　网址：www.tzmcm.cn
电话：0471-4969085　　邮编：010021　　Email：2014@tzmcm.cn

2014 年第七届“认证杯”数学中国数学建模网络挑战赛第一阶段论文

题　　目　土地储备方案的风险评估

关 键 词　决策树　C4.5　土地储备贷款　风险度量　数据挖掘　K—均值聚类　因子分析　logistic 回归

摘　　要：

本文主要运用数据挖掘中的决策树技术对经过预处理后的土地储备贷款挖掘数据集中的数据进行分析，发现隐藏在大量数据中的隐含模式，最终得到土地储备贷款风险评估模型。决策树算法作为数据挖掘体系中的一项重要分类预测方法，有着简单、高效、结构清晰等优点。本文研究的内容主要包括数据采集、数据预处理以及模型的建立，而这三部分又是基于数据挖掘以及决策树的相关理论，根据数据挖掘工作的一般步骤展开的，论文针对在金融衍生产品还未大量涉及土地相关资产的客观情况下进行的土地储备风险管理。首先建立土地储备风险度量模型，借鉴外资银行应用数据挖掘决策树技术建立客户信用评价系统的成功经验，在明确挖掘目的的前提下，深入理解数据挖掘、决策树、数据采集、数据预处理、聚类分析以及模型评价等方面。在此基础上，结合本文研究的问题以及数据的特点，通过各种方法的分析与比较，应用决策树 C4.5 算法以及聚类 k 一平均算法对这些数据进行挖掘和分析，通过计算风险度量值和决策影响程度值，确每个属性各个取值的分数值，得到土地储备贷款风险评估模型。我们将 1、2、23、47、48、49、54、66、72、74 这 10 个风险最大的项目提供给土地储备部门（理由见正文）。土地风险储备评估是土地收储过程的一个重要环节，其预测精度和有效性直接关系到土地收储部门和信贷机构的损益以及金融市场的繁荣。该评估模型在实际应用中对于土地储备中心的风控人员做贷款决策具有一定的指导性作用，能够为土地储备风险控制提供支持。

参赛队号：　2351
所选题目：　C　题

参赛密码 _____
（由组委会填写）

英文摘要（选填）

第七届数学中国数学建模网络挑战赛
地址：数学中国数学建模网络挑战赛组委会 网址：www.tzmcm.cn
电话：0471-4969085 邮编：010021 Email：2014@tzmcm.cn

In this paper, we use the decision tree data mining techniques to analyze data set of attachment, we did data pre-processing jobs before analyzing, and we find out mode which is behind the massive data set. Finally, we conclude the risk of land reserve loans assessment model. Decision tree algorithm is a important method of classification and prediction, it's easy, efficient and explicit. This paper mainly talk about the data collection, data preprocessing and how to build the model, and those parts are based on data mining, decision tree theory and the basic step of data mining process. This paper emphasize that financial derivatives has not been heavily involved in land-related assets of the objective situation of the land reserve risk management. Firstly, we build the measure modeling of land risk reserve, and import successful experience of bank, under the premise of a clear purpose to deeply understand the data mining, decision tree, data acquisition, data preprocessing, clustering analysis, and model evaluation. On this basis, combining this problem and the characters of data, compare with a lot of method, and use the C4.5 decision tree algorithm and k average clustering method to data mining and analysis those data, By calculate the degree of risk measures and decisions affect the value to determine the score of every attribute, and conclude the risk of land reserve loan assessment modeling. Finally we found that 1，2，23，47，48，49，54，66，72，74 risk is high. The land reserve risk assessment is a important part of land reserve process, the accuracy prediction and effectiveness is directly related to profit or loss as well as the prosperity of the financial market land purchasing and storage departments and credit institutions. The combined model in practical applications for risk control personnel land reserve center to make some decisions loans guiding role, providing support for the land bank risk control

一、问题重述

土地储备风险的影响因素及其影响程度分析：风险的基本含义是指损失的不确定性。具体而言，土地储备风险是指城市土地储备制度的建立和运营过程由于各种事先无法预料的不确定因素带来的影响，使得土地储备的实际收益与预期收益发生一定偏差，从而有蒙受损失的机会或可能性。我国现行的土地制度，与西方国家的土地制度存在着很大差别。这不仅决定了我国土地市场运作机制的特殊性，也决定了我国土地储备中风险存在形式的独特性。因此，城市中土地储备风险的正确测度与有效的规避，是土地储备工作面临的重要问题。

土地储备风险的影响因素分析：随着城市土地储备机构在我国大多数省市的建立运营，城市土地储备制度已经显示出其旺盛的生命力，产生了巨大的经济社会效益，如在城市的旧城改造、企业改制、腾笼置业等方面取得了一些成绩。然而，由于这是一种制度创新，土地储备制度还处于发展阶段，在其运行中面临着诸多风险。主要有金融风险具体如下：

资金和土地是城市土地储备制度得以建立并运营的两大基本要素。城市土地储备的过程不仅是土地流转的过程，更是资金循环周转的过程，因此资金落实与否是土地储备能否达到预期目的的重要因素，如果没有大量的资金加以支撑，城市土地储备就难以运作。土地储备属于资金密集型的业务，当前全国各地土地收购储备中心在资金筹措上大部分依赖银行贷款，而且银行贷款所占比例相当高。以土地储备较成熟的浙江省为例，2003 年 3 月份的统计数字显示，浙江全省土地收购储备资金达 112 亿元，其中 100 亿元是银行贷款，占总量的 89%，一些地级市土地储备资金中银行贷款比重高达 95%(陈平，2003)。由此可见，土地储备资金的来源渠道和资金运行风险就成为城市土地储备工作中的主要问题。

土地储备中心自有资本金短缺，资金补充机制不健全土地储备中心大多为事业编制的事业法人，中心运作所需的启动资金一般是由财政注入一定的资本金作为注册资金或临时启动资金(于水，2002)。从整体而言，各地注入土地储备运作的资金极其有限，甚至，少数城市因财政拿不出资金，而由政府划拨一定的土地作为中心的启动“资金”。中心再以划拨土地作为抵押向银行申请贷款进行运作，而无实际资金投入。土地储备机构由于自有资金短缺，资金补充机制不完善，只好依赖银行大规模贷款运作，而在贷款运作中自有资金与信贷资金的比率偏低，据《浙江省土地储备工作的调研报告》显示，土地储备机构最高的自有资金占有率仅为 15%，根据市场规律，一个企业的自有资金占所运营的资金的 30%时，属于安全运作。此外银行贷款资金虽一般能及时到位，但利息压力太大，使土地储备机构债务负担沉重，据调查，有的城市土地储备中心一天的银行利息即达数十万元。更为重要的是，从现代企业运营来看，单一的融资渠道和畸高的债务(即贷款比例高，期限短，与收购储备资金运用存在时间上、期限上的不匹配)是企业运营的危险地带。

土地储备机构的利率风险至关重要由于土地储备资金主要来源于银行贷款，过高的贷款比例给土地储备中心带来了沉重的利息负担，造成了储备资金融资渠道狭窄，且成本高，风险大。不仅难以满足城市土地储备机构在收购城市土地时对资金的大量需求，而且可能会因为银行利息增加而使储备土地的成本大幅增加，土地储备中心有时迫于偿还银行本息的压力，在短期内过量地向市场供地，使得储备的土地不能在合适的时机和价位出手，从而造成收入的不必要减少，甚至还可能出现城市土地出让收益低于储备土

地成本的情况，使得土地储备计划难以完成，这与政府设立土地储备中心的初衷相违背(刘新芝，2005)。

土地储备机构的财务风险巨大来自财务方面的风险因素主要有两个：融资风险和违约风险。融资风险是指在储备土地时运用财务杠杆，即使用贷款的条件下，虽扩大了投资的利润范围，但也增加了不确定性，新增营业收入不足以偿还债务的可能性。财务杠杆受预期利润率和贷款利率两方面的影响，投资的利润范围扩大，投资的不确定性增大，不能达到预期收益的可能性也随之增大。这是因为，财务杠杆的使用提高了税前年收益的期望值和可能收益的上限，但是也扩大了年收益波动的范围，降低了可能收益的下限，下端风险增大了，加上抵押贷款贷方对净收入有优先要求权，而投资者的税前现金流量是还贷后的余额，所以增加贷款量的同时也增加了营业收入不足以偿还债务的可能性。违约风险是指储备土地出让或出租过程中，由于受让者或承租者财务状况恶化而使土地投资及其报酬无法全部收回的可能性，或者是受让者或承租者不按约按期支付款项，拖欠严重，使投资者入不敷出所造成的一种风险。

最后，出让风险也是重要影响因素之一，它是指土地出租、出让交易过程中由于各种不确定因素带来的风险。一方面是土地交易“流拍”，使部分地块不能实现顺利转让，结果会导致地价低于成本出售或土地积压。另一方面是片面强调价高者得，引发不理性竞争或出现资源配置的错位。

二、问题分析

针对问题 1，我们对某省级土地储备中心的土地储备项目可研报告的数据进行数据挖掘，建立合理的土地储备风险评估模型，为土地储备部门提供一个比较实用的土地储备方案的风险评估方法。

针对问题 2 ，我们对土地储备项目可研报告有人为修改的情况，利用我们的风险评估方法对附件二中的方案进行风险评估，将 10 个风险最大的项目提供给土地储备部门，并从数据挖掘和风险度量模型的角度，指出造成这 10 个项目风险较大的原因。

我们研究的内容主要包括数据采集、数据预处理以及模型的建立，而这三部分又是基于数据挖掘以及决策树的相关理论，根据数据挖掘工作的一般步骤展开的，论文针对在金融衍生产品还未大量涉及土地相关资产的客观情况下进行的土地储备风险管理。首先建立土地储备风险度量模型，借鉴外资银行应用数据挖掘决策树技术建立客户信用评价系统的成功经验，在明确挖掘目的的前提下，深入理解数据挖掘、决策树、数据采集、数据预处理、聚类分析以及模型评价等方面的相关概念。在此基础上，结合本文研究的问题以及数据的特点，通过各种方法的分析与比较，用适合的数据采集和预处理方法对数据库中的一半数据进行处理，建立适合挖掘的数据集，应用决策树 C4.5 算法以及聚类 k 一平均算法对这些数据进行挖掘和分析，通过计算风险度量值和决策影响程度值，确每个属性各个取值的分数值，得到土地储备贷款风险评估模型。最后使用剩余的一部分数据作为测试样本来评价这个模型，证明此模型具有较强的预测能力，是当前土地储备中心可以采用的最优模型，值得在实践中推广。

2.1 数据挖掘的定义

1995 年，在美国计算机年会(ACM)上，提出了数据挖掘的概念，到目前为止，数据挖掘的定义很多，本文从技术和商业两个角度分别进行阐述。

2.1.1 技术角度的定义

从技术角度看,数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的，人们事先不知道的，但又是潜在有用的信息和知识的过程。与数据挖掘相近的同义词有数据融合，数据分析和决策支持等。这个定义包括好几层含义:数据源必须是真实的、大量的、含噪声的，发现的是用户感兴趣的知识，发现的知识要可接受、可理解、可运用，并不要求发现放之四海皆准的知识,仅支持特定的发现问题。

2.1.2 商业角度的定义

数据挖掘是一种新的商业信息处理技术,其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理,从中提取辅助商业决策的关键性数据。目前,由于各行业业务自动化的实现,商业领域产生了大量的业务数据,这些数据不再是为了分析的目的而收集的,而是由于纯机会的商业运作而产生″分析这些数据也不再是单纯为了研究的需要,更主要是为商业决策提供真正有价值的信息,进而获得利润。但所有企业面临的一个共同问题是:企业数据量非常大,而其中真正有价值的信息却很少。因此从大量的数据中经过深层分析,获得有利于商业运作，提高竞争力的信息,就像从矿石中淘金一样,数据挖掘也因此而得名。因此,数据挖掘可以描述为:按企业既定业务目标,对大量的企业数据进行探索和分析,揭示隐藏的、未知的或验证已知的规律性,并进一步将其模型化的先进有效的方法。

2.2 数据挖掘的作用及功能

对于企业而言,数据挖掘可以有助于发现业务发展的趋势,揭示已知的事实,预测未知的结果,并帮助企业分析出完成任务所需的关键因素,以达到增加收入、降低成本,使企业处于更有利的竞争位置的目的。

数据挖掘通过预测未来的趋势及行为,可以让企业做出前缀的、基于知识的决策。数据挖掘的目标是从数据库中发现隐含的、有意义的知识、主要有以下 5 类功能：

1.自动预测行为和趋势

利用历史数据找出规律,建立模型,并用此模型来预测未来数据的种类特征等。一个典型的例子是市场预测问题,数据挖掘使用过去有关促销的数据来寻找未来投资中回报最大的用户，其他可预测的问题包括预测破产以及认定对指定事件最可能做出反应的群体。

2.关联分析

数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性,就称之为关联″关联可分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网，有时并不知道数据库中数据的关联函数,即使知道也是不确定的,因此关联分析生成的规则带有可信度。

3.聚类

数据库中的记录可被划分为一系列有意义的子集,称为聚类。聚类增强了人们对客观现实的认识,是概念描述和偏差分析的先决条件。聚类技术主要包括传统的模式识别方法和数学分类方法。又十年代初,Mchalski 提出了概念聚类技术的要点是:在划分对象时不仅考虑对象之间的距离,还要求对划分出的类具有的某种内涵进行描述,从而避免了传统技术的某些片面性。

4.概念描述

概念描述就是对某类对象的内涵进行描述,并概括这类对象的有关特征概念描述分为特征性描述和区别性描述,前者描述某类对象的共同特征,后者描述不同类对之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性。生成区别性描述的

方法很多，如决策树方法、遗传算法等。

5. 偏差检测

数据库中的数据常有一些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是：寻找观测结果与参照值之间有意义的差别。

2.3 数据挖掘的一般过程及具体应用

2.3.1 数据挖掘的一般过程

挖掘过程一般需要经历确定挖掘对象、准备数据、建立模型、数据挖掘、结果分析与知识应用这几个阶段，图 2 一 1 描述了数据挖掘的基本过程和主要步骤：

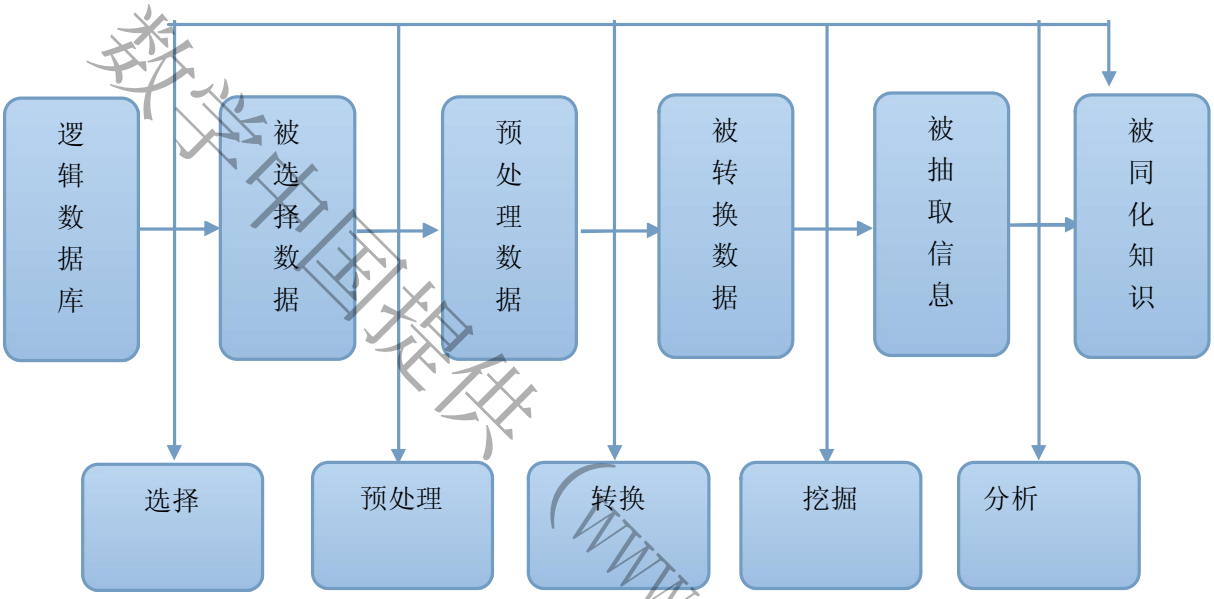


图 2 一 1 数据挖掘流程

过程中各步骤的大体内容如下：

1. 确定挖掘对象

清晰地定义挖掘对象以及挖掘目的是数据挖掘的重要一步，挖掘的最后结构是不可预测的，但要探索的问题应是有预见的。

2. 准备数据

(1) 数据选择

搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘的数据，集成和合并数据到单一的数据挖掘库中，并协调来自多个数据源的数据在数值上的差异。

(2) 数据预处理

选择数据后，还需要对数据进行预处理，对数据进行清洗，解决数据中的缺值、冗余、数据值的不一致、过时的数据等问题。

3. 建立模型

建立一个真正适合挖掘算法的分析模型，是数据挖掘成功的关键。模型的建立必须从数据的分析开始，首先为模型选择变量。接着，从原始数据中构建新的预示值"下一步，就需要从数据中选取一个子集或样本来建立模型。最后，需要转换变量，使之和选定用来建立模型的算法一致。

4. 数据挖掘

对所得到的经过转化的数据进行挖掘,除了完善与选择合适的算法需要人工干预外,数据挖掘工作都由挖掘工具自动完成。

5.结果分析

当数据挖掘出现结果后,要对挖掘结果进行解释并且评估。其使用的分析方法一般应视数据挖掘操作而定。

6.知识应用

为将数据挖掘结果能在实际中得到应用,需要将分析得到的知识集成到业务信息系统的组织结构中去,使这些知识在实际的管理决策分析中得到应用。

2.3.2 土地储备贷款数据挖掘的过程

本文研究的主要目标是通过数据挖掘工具对大量信息进行处理分析,以求达到对贷款进行风险评估和防范土地储备贷款风险,尽量减少风险贷款的发生的目的。具体来说,本研究要实现的目标包括:

1.建立土地储备贷款评分模型:通过对数据的挖掘,对土地储备方案的不同类别属性给予不同分值,得到可以量化的土地储备贷款评分体系,从而可以从挑选土地贷款方案的角度防止金融风险的发生。

2.得到土地储备贷款等级分类规则:根据评分体系得到土地储备贷款评分,进一步得到对不同分值土地储备贷款等级的分类规则。

3.预测土地储备贷款方案情况:通过样本数据的训练提取关于这些数据的特征式,当新的土地储备贷款情况输入时可以较为准确的确定该土地储备贷款等级。

4.建立风险评估评分模型:根据土地储备贷款状况与贷款的其它特征综合评价风险等级。

为了达到这个目标,首先根据土地储备方案的贷款信息,采集土地储备项目可研报告中的数据,经过数据预处理得到可以进行数据挖掘的待处理数据集。然后,通过模型进行贷款风险等级分类,预测贷款风险等级,得到决策模型。经过知识评价、结论解释和知识提取,形成知识库(知识库也将进一步提高模型库预测、分类以及决策的能力)。最后将经过结论解释得到的信息作用于土地储备贷款风险评估,将作用后得到的反馈信息作用于进一步的风险控制需求中,并根据评估适当修改采集的数据集。

2.4 数据挖掘的方法与选择

我们研究的土地储备贷款信用风险评估的问题属于分类问题。而对于分类问题,分类方法主要有三种:基于传统统计分析的数据分类方法、基于神经网络的数据分类方法和基于决策树技术的数据分类方法。

2.4.1 传统统计分析统计分析是数据挖掘算法中最基础的部分。

许多数据挖掘技术都利用了存在已久的统计技术。这类技术包括相关分析、回归分析及因子分析等， 多元统计分析包括因子分析、聚类分析等。

统计预测方法包括回归分析、时间序列分析等。例如:抽样技术面对的是大量的数据,对所有的数据进行分析是不可能的,也是没有必要的,因此就要在理论的指导下进行合理的抽样。在这些分析过程中,一般先由用户提供假设,再由系统利用数据进行验证。此外,统计分析在辨别分析和回归建模方面有着自己独特的长处。辨别分析在对于客户价值细分方面很有作用,在回归建模方面,广泛地用于预测顾客将来的行为,例如预测客户的潜在价值和未来的购买愿望等。

统计方法的最大优点在于其具有明显的解释性,存在的缺陷是过于严格的前提条件（样本量少、正态分布、等协方差等)。也就是可解释性比较好,速度快,但是由于样本要求是样本量少并且要求数据的完整性,所以针对海量数据和不十分完善的数据,基于传统统计方法的模型其预测的准确率、强壮性和可伸缩性都比较差。

2.4.2 神经网络技术

神经网络为人们解决大复杂度的问题提供了一种有效的简单方法,它可以很容易的解决具有上百个参数的问题,主要应用于分类和回归两类问题.在结构上,神经网络划分为输入层、输出层和隐含层,如图 2 一 2 所示:

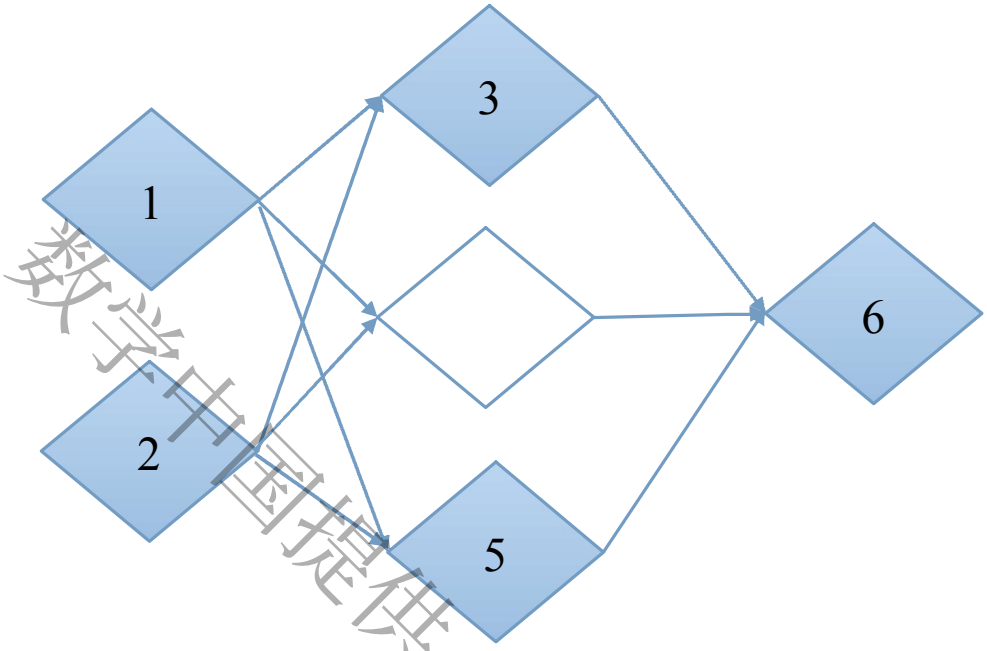


图 2 一 2 神经网络示意图

输入层每个节点对应一个个的预测变量。输出层节点对应目标变量,可有多个"在输入层和输出层之间是隐含层,隐含层层数和每层节点的个数决定了神经网络的复杂度。除了输入层的节点,神经网络的每个节点都与很多它前面的节点连接在一起,每个连接对应一个权重 Wxy,此节点的值就是通过它所有输入节点的值与对应连接权重乘积的和作为一个函数的输入而得到,我们把这个函数称为活函数或挤压函数。

神经网络的每个节点都可表示成预测变量的值或者值的组合。节点 6 的值已经不再是节点 1、2 的线形组合,因为数据在隐含层中传递时使用了活动函数。调整节点间连接的权重就是在建立神经网络时要做的工作。神经元网络和统计方法在本质上有很多差别。神经网络的参数可以比统计方法多很多。如上图中就有 13 个参数(9 个权重和 3 个限制条件),神经元网络的优势是运行分析时无需心中有任何特定模型,而且通过神经网络可以发现交互作用的效果(如年龄和性别的组合效果)。神经网络的缺点是不易用它的权重层和晦涩的转化来解释结果模型。神经网络在数据为高度非线形并有交互作用时对预示目标变量非常有用,但在需要解释数据中的关系时就不太有帮助。

2.4.3 决策树方法

决策树方法是一种简单的知识表示方法,一般用于分类任务,它将事例逐步分类,代表不同的类别。由于分类规则是比较直观的,因而此法比较容易理解,决策树类似于流程图中的树结构,其中每个内部节点表示在一个属性上的测试,每个分枝代表一个测试输出,而每个树叶节点代表类或类的分布。决策树提供了一种类似在什么条件下会得到什么值这类规则的方法。比如,在贷款申请中,要对申请的风险做出判断"决策树的基本组成部分是:决策节点、分支和叶子。决策树中最上面的节点称为根节点,是整个决策树的开始。决策树中的每个节点的下层子节点的个数与决策树使用的算法有关。如 CART 算法得到的决策树的每个节点有两个分支,这种树就称为二叉树,允许节点含有多于子

节点的树称为叉树。每个分支要么是一个新的决策节点,要么是树的结尾,称为叶子。在沿着决策树从上到下遍历的过程中,在每个节点都会遇到一个问题,对问题的不同回答导致不同的分支,最后会到达一个叶子节点。这个过程就是利用决策树进行分类的过程,利用几个变量(每个变量对应一个问题)来判断所属的类别(最后每个叶子会对应一个类别)。

决策树与神经元网络相比较,其优点在于可以生成一些规则。当我们进行一些决策,同时需要相应的理由的时候,神经元网络就不可行。决策树是分析消耗(流线型生产)、发现交叉销售机会、进行促销、信用风险或破产分析和发觉欺诈行为的得力工具。决策树方法的优点在于:可以生成可以理解的规则，计算量相对来说不是很大，可以处理连续和种类字段，可以清晰的显示哪些字段比较重要；缺点在于:对连续性的字段比较难预测，对有时间顺序的数据需要很多预处理的工作，当类别太多时,错误可能就会增加的比较快，一般算法分类的时候,只是根据一个字段来分类。

2.4.4 数据挖掘方法的选择

在选择数据挖掘分类方法时,我们需要考虑以下几个方面:

1.预测的准确率,即模型正确地预测新的或先前未见过的数据的类标号的能力。

2.速度,即产生和使用模型的计算花费。

3.强壮性,即给定噪声数据或具有空缺值的数据,模型正确预测的能力。

4.可伸缩性,即给定大量数据,有效地构造模型的能力。

5.可解释性,即学习模型提供的理解和洞察的层次。

对比这三种分类方法,结果列表 2—1 所示:

表 2—1　分类方法比较

比较内容	传统统计方法	神经网络技术	决策树方法
预测的准确率	一般	较好	较好
速度	较快	较慢	较快
强壮性	差	一般	一般
可伸缩性	差	一般	好
可解释性	强	弱	强

通过对三种分类方法的对比可以发现,决策树方法有几个显著的优点:

1.能够生成可以理解的规则。

2.计算量相对来说不是很大,所以计算速度较快。

3.可以处理连续和离散的字段。

4.可以清晰地显示哪些字段比较重要。

5.训练精度高。

6.决策树很擅长处理非数值型数据。

结合决策树方法的优点,综合考虑本文研究的目标要求和数据特点,本文决定选用决策树分类方法作为土地储备贷款风险分类的方法。

2.5 决策树的基本理论

2.5.1 决策树的基本概念

决策树因其形状像树且能用于决策而得名,因其出色的数据分析效率、直观易懂的特点备受青睐,成为数据挖掘常用的技术。它通过将大量数据有目的地分类,从中找出一些潜在的、对决策有价值的信息。从技术上讲,一个决策树由一系列结点和分枝组成,树中的每个非叶结点(包括根结点)对应于训练集中一个非类别属性的测试,非叶结点的每一个分枝对应属性的一个测试结果,每个叶子结点则代表一个类或类分布。从根结点到

叶子结点的一条路径形成一条分类规则,决策树可以很方便地转化为分类规则,是一种非常直观的分类模式表示形式。

在金融风险领域常用决策树来分析数据并作出结论.例如,银行在个贷业务中,可先对客户贷款风险的高低进行评估。决策树的构建是一种自上而下、分而治之的归纳过程。其中,测试属性的选择和如何划分样本集是构建决策树的关键环节。不同的决策树算法在此使用的技术也不尽相同。

2.5.2 决策树算法及选择

国际上最早和最有影响的决策树方法是由 Quiulan 研制的 ID3 决策树生成算法,该算法采用信息增益作为属性选择的度量标准。后来,他又提出了改进版本 C4.5 算法。该算法的基本工作流与 ID3 算法相同,但 C4.5 算法采用信息增益率作为属性选择的度量标准,还增加了对连续属性的离散化，对未知属性的处理和产生规则等功能。下面分别介绍这两种算法,并选择适合的方法作为本研究决策树的构造算法。

2.5.3 ID3 算法

在决策树构造中,ID3 算法通常采用信息增益方法确定根据哪个属性来产生分支。设 S 是 s 个数据样本的集合,假定类标号属性具有 m 个不同类 $C_i(i=1,\cdots,m)$,设 s_i 是类 C_i 中的样本数。对一个给定的样本分类所需的期望信息由式(3.1)给出:

$$I(S_1,S_2,\ldots,S_m)=-\sum_{i=1}^{m}p_i\log 2(p_i) \tag{3.1}$$

p_i 是任意样本属于 C_i 的概率，$p_i=S_i/S$

设置属性 A 具有 V 个不同的值 $\{a_1,a_2,\ldots,a_v\}$,可以用属性 A 将 S 划分为 V 个子集 $\{s_1,s_2,\ldots,s_v\}$。其中，包含 S_j 中有这些样本，它们在 A 上具有值 a，如果 A 选作测试属性(既最分裂属性),则这些子集对应于由包含集合 S 的节点生长出来的分枝。设 S_{ij} 是 S_j 中类 C_i 的样本数，根据有 A 划分成子集的熵（emtropy）或期望信息由式（3.2）给出：

$$E(A)=\sum_{i=1}^{v}\frac{s_{1j}+\ldots s_{mj}}{s}I(s_{1j},\ldots,s_{mj}) \tag{3.2}$$

其中项 $\frac{s_{1j}+\ldots s_{mj}}{s}$ 充当第 j 个子集的权，并且等于子集（既 A 值为 a_j）中的样本个数除以 S 中样本总数。熵值越小，子集划分的纯度就高。对于给定的子集 s_j：

$$I(S_1,S_2,\ldots,S_m)=-\sum_{i=1}^{m}p_i\log 2(p_{ij}) \tag{3.3}$$

其中，$p_{ij}=s_{ij}/s_j$ 是 s_j 中的样本属于类 C_i 的概率。

在 A 上划分获得的信息增益为：

$$Gain(A) = I(s_1, s_2, \cdots, s_m)E(A) \quad (3.4)$$

换言之，$Gain(A)$ 是由于获得属性 A 的值而导致的熵的期望压缩。通过计算每个属性的信息，然后选择增益最大的那个属性作为给定集合 S 的测试属性，并由此产生相应的分支结点。ID3 算法是最经典的决策树算法，应用非常广泛，但它存在着很多不足。

1. 存在种类偏见问题，即信息增益的计算倾向于选择取值较多的属性，但取值多的属性不一定是最优的。

2. ID3 算法构造的决策树是单变量决策树，忽略了属性间的相互联系。

3. ID3 算法不能直接处理连续性属性。

4. 不能处理属性值空缺的样本。

正是因为 ID3 算法存在以上不足，Quialan 于 1993 年又研制了 C4.5 算法。

2.5.4 C4.5 算法

C4.5 算法是以 ID3 算法为核心的完整的决策树生成系统，它通过树的生成和树的剪枝这两个步骤来建立决策树。与 ID3 算法不同的是，C4.5 选取使得信息增益率最大的属性作为测试属性"信息增益率等于信息增益对分割信息量的比值，定义如下：

$$Ratio(A) = Gain(A) / E(A) \quad (3.5)$$

该算法需要计算每个决策属性的信息增益率，具有最高信息增益率的属性被选作给定集合 S 的测试属性，创建一个节点，并以该属性标记，对属性的每个值创建分枝，并且据此划分样本。对于连续值属性 A，C4.5 按照属性的信息增益率将其划分为两个不同的子集：属性值大于分割点和属性值小于等于分割点，即使用如下的测试形式：$A < r$ 和 $A > r$，r 为分割点

C4.5 寻找最优的 r 的方法是：

1. 首先采用快速排序法将训练集的样本根据属性 A 的值排序.

2. 然后按顺序逐一将两个相邻的样本的 A 的平均值，$r = (A_1 + A_2)/2$ 作为分割点（假设训练集有 n 个样本，则共有 $n-1$ 个分割点）每个分割点都可将训练集划分为两个子集，划分后所得的信息增益 $Gain_r$，线性扫描 $r_1, r_2, \cdots r_{n-1}$，比较所有可能的分割点，使得 $Gain_r$ 最大并将其作为最优的分割点 r。按照上述方法求出当前侯选属性集所有属性的信息增益率，选出信息增益率最高的属性，然后按照该属性的分割点，将当前样本分为两个子样本集。对子样本集采用同样的方法继续分割直到不能再分割或达到停止条件为止。

C4.5 处理空缺属性值的方法是：在计算系统整体的不确定时，只根据那些已知测试属性值的样本来计算。在计算根据某属性划分后的信息熵时，把那些丢失测试属性值的样本作为一个新的子样本集，单独计算这些样本的期望信息熵。在划分训练集时，先将不

含空缺属性值的样本按照一定的算法划分为几个子集，然后把那些丢失测试属性值的样本按照一定的概率分布到各个子集中，子集中含有空缺测试属性值的样本与有测试属性值的样本保持一个比例关系。在对含有空缺测试属性值的未知实例进行分类时，C4.5 将该实例通过所有的分支，然后将结果进行合并，使它成为在类上的概率分布而不是某一个类，然后选择具有最大概率的类作为最终的分类结果。

C4.5 算法生成决策树之后，还将进行树剪枝。对每个叶子结点，分类错误率为该结点中不属于该结点所表示类别的样本的权值之和。对非叶结点，分类错误率为它的各个子结点的分类错误率之和。如果计算出某结点的分类错误率超过了将结点 N 所代表的样本集 T 中的所有样本分配为 T 中出现最多的类别所得的分类错误，则将结点 N 的所有子枝剪去，使 N 成为叶结点，将 T 中出现最多的类别分配给它。

由于C4.5采用训练样本集来估计分类错误率，因此会使得到的决策树融进了训练集中的某些异常，而这些异常通常在总体样本中并不出现，从而导致决策树倾向于过度拟合(Overfitting)这个缺陷可以使用一种悲观估计来补偿，即选择一组独立于训练样本集的测试样本集来优化决策树。

2.5.5 算法比较与选择

虽然C4.5算法在选择测试属性，分割样本集上采用的技术仍然没有脱离信息熵原理，但与 ID3 算法相比，C4.5 算法在效率上有了很大的提高，不仅可以直接处理连续型属性，还可以允许训练样本集中出现属性空缺的样本，生成的决策树的分枝也较少。从理论上讲，C4.5 决策树算法较完善，且简单易懂，生成速度也比较快，通过生成的决策树，可以生成可理解的规则。

因此，本文引用了C4.5算法对申请个人住房贷款的信用数据进行实证分析建立贷款人是否违约的分类模型，应用于银行个人住房贷款信用风险的评估与预测。

2.5.6 决策树的构造

决策树生成的操作过程：

决策树生成的操作过程如图 2 一 3 所示：

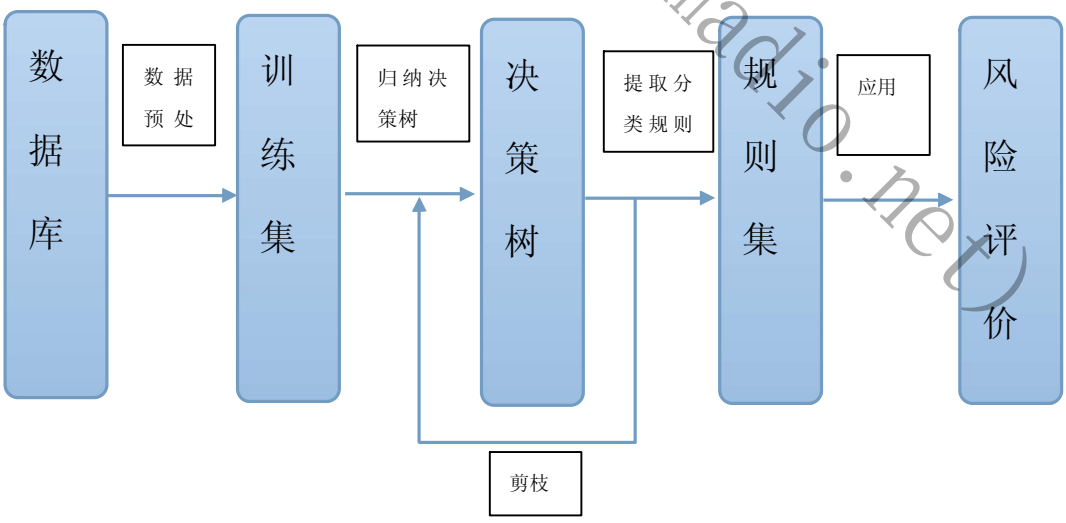


图 2 一 3 决策树生成操作过程

如图 2 一 3 决策树生成操作过程

1. 系统从内部网各接触点收集土地储备信息，对数据信息进行合并，形成结构统一的土地储备信息数据源。

2.对数据源进行数据预处理,去掉与决策无关的属性和高分支属性，将数值型属性进行概化以及处理含空缺值的属性,形成决策树的训练集。

3.对训练集进行训练,计算每个属性的信息增益率,选择信息增益率最大的属性作为当前的主属性节点,为该属性的每一个可能的取值构建一个分支。对子结点所包含的样本子集递归地执行上述过程,直到子集中的数据记录在主属性上取值都相同,或没有属性可再供划分使用,生成初始的决策树。

4.对初始决策树进行树剪枝，主要采用后剪枝算法对生成的初始决策树进行剪枝,并在剪枝过程中使用一种悲观估计来补偿树生成时的乐观偏差。

5.由所得的决策树提取分类规则，对从根到树叶的每一条路径创建一个规则,形成规则集。

6.系统运用决策树所得规则对新数据进行分析,预测该数据的类别,帮助银行进行决策。

决策树构造的输入是一组带有类别标记的数据集,构造的结果是一棵二叉或多叉树。二叉树的内部节点(非叶子节点)一般表示为一个逻辑判断,如形式为 $a_i = v_i$ 的逻辑判断,其中 a_i 是属性, v_i 是属性值，树的边是逻辑判断的分支结果。多叉树的内部节点是最佳扩展属性,叶节点是类别属性值。内部节点的射出边是最佳扩展属性的取值,有几个属性值,就有几条边。内部节点对应的数据集是不纯的(数据属于多种类别),根节点对应的数据集是训练集 D,其它内部节点对应的数据集是 D 的子集,叶节点对应的数据集是纯的训练子集(数据属于同一类),树的叶子节点都是类别标记。构造决策树分为两步:

1.决策树的生长:由训练集生成一棵决策树。

2.剪枝:用非训练集中的事例检验生成的决策树,剪去影响预测精度的分枝。

2.5.7 决策树的生长

创建一棵决策树可以递归地实现。首先,使用之前介绍的知识计算各属性的信息增益率,选择信息增益率最大的属性作为根节点,然后把该属性的每一个可能的值作为子节点,这样就把整个数据集分成了几个子集。根节点属性的每个值都是一个子集,现在这个过程可以递归地应用到每个子树上进行进一步的划分,在任何时候,如果子集中的所有元素都是同一类的,则停止划分。

C4.5 算法核心部分的描述如下:

```
/*参数:R 表示一些带有连续取值的非目标属性,C 表示目标属性,S 表
示训练集*/
FunetionC4.5(R，C，S)
{
For(R 中的属性 Ri)
{
If（Ri 的取值是连续的）
{

A1=min（Ri）;//A1为 R1的最小值

Am=max(Ri)://Am 为 Ri 的最大值
For （j=2;j<m； j++//m 的值是手工设定的
```

Aj=A1+j*(Am 一 A1)/m;

得到基于{<=A^j，>A^j}分类的最大增益 Gain（R^i，S)；

}

D 为属性集 R 中最大增益 Gain(D,S)的属性;

{d^j｜j=1，2,…m}为属性 D 的取值;

{S^j｜j=1,2,…m}为与 S 相对应的包含属性 D 相应取值 dj 的数据集;

//递归得到包含属性 D 相应取值鸣的各个子决策树

C4.5(R2{D},C,S1),C4.5(R2{D},C,S2),.二,C4.5(R2{D},C,Sm);

返回一个根为 D 的树;

}

2.5.8 C4.5 算法决策树的修剪

现实世界中的数据一般有缺值,不完整，不准确和噪声等，剪枝是一种客服噪声的技术,同时也能使决策树得到简化,变得更加容易理解"得到了完全生长的初始决策树后,为了除去噪声数据和孤立点引起的分枝异常,C4.5 采用后剪枝算法对生成的初始决策树进行剪枝。决策树的剪枝通常是用叶结点替代一个或多个子树,然后选择出现概率最高的类作为该结点的类别,在 C4.5 中还允许用其中的树枝来替代子树。

一个覆盖 N 个实例,其中 E 个为错误的。对于具有 CF (聚类特征)的实例,计算一个二项分布 $B_{cf}(E,N)$，该二项分布为实例的误判概率。N 个实例判断错误数为 $N*U_{cf}(E,N)$。子树的错误数为所有叶结点的总和。如果使用叶结点或者树枝代替原来的子树之后,误差率若能够下降,则使用此叶结点或者树枝代替原来的子树。

2.5.9 C4.5 算法规则提取

对于生成的决策树,可以直接从中提取规则"过程是将决策树转化成比较直观的规则形式,更好地理解分类结果。分类规则是用 IF 一 THEN 形式表示,每条规则都是一条从根到叶节点的路径,叶结点表示具体的结论,而叶结点以上的结点及其边表示的相应条件的条件取值。从决策树到决策规则的转换如图 3 一 3 所示:

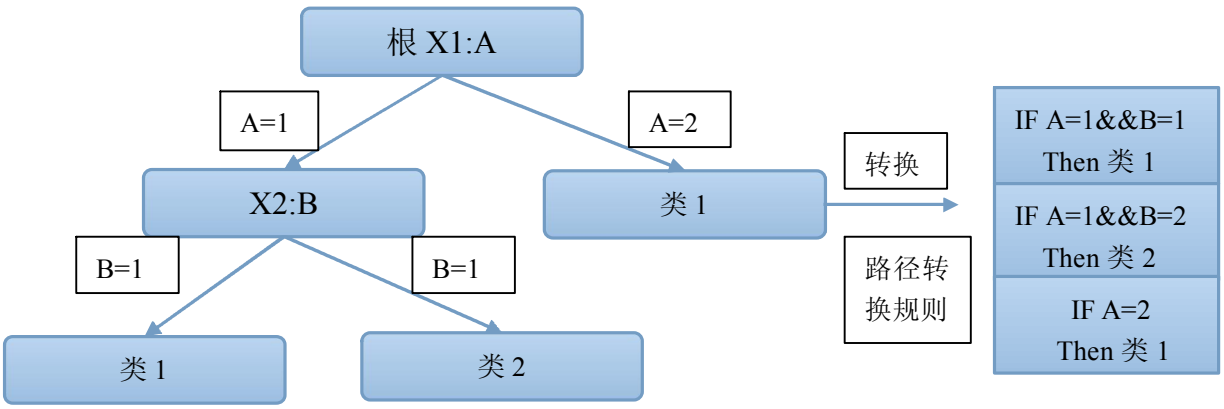


图 2 一 4 决策树到决策规则的转换

总之,C4.5 算法是一种适用范围比较广泛、率较高的决策树算法"通过数据预处理,参数和类选定,构造和修剪决策树,进行分析和评估,生成分类规则等步骤后,完成分类挖掘。后文将运用 C4.5 算法对土地储备贷款方案数据进行处理、挖掘和分析。

三、模型一的建立求解与结论分析

3.1 数据降维去噪过程及结果

3.1.1 数据准备

前面的内容中我们已经介绍了本研究的理论基础,现在开始,我们将把理论应用于实际,具体介绍土地储备贷款风险评估模型的构造过程。而模型的构建必须建立在数据之上,这部分我们通过数据采集及数据预处理,得到适合挖掘的土地储备贷款挖掘数据集。

3.1.2 数据采集

数据采集方法在问题定义完毕之后,就可以根据土地储备项目可研报告去获取原始数据,建立相应的挖掘数据库,也就是要为土地储备贷款的挖掘任务准备数据。在本研究中,需要的数据存在于土地储备项目可研报告中,但是,土地储备项目可研报告的数据存在各种类型的数据表和大量的数据字段,数据挖掘算法也无法从中获取结果。因此,需要从土地储备项目可研报告中提取数据,并将它们存储为数据挖掘算法可以处理的模式。对于土地储备项目可研报告来说,数据表很多,结合本文研究的问题,经过分析得到 5 个与本研究有关的数据表:拟收储地块情况统计表、交易案例情况调查表、土地供应情况调查表、土地一级市场供应情况调查表、近年土地收储情况调查表。然后,抽取各表中对挖掘有用的信息,对表进行并运算,使之成为一张普通的二维表。另外,对相同的记录进行合并和计数。

本研究的数据库中存在3922条个土地储备贷款数据,用这样的数据量去进行数据挖掘显然是不合适的。我们使用随机序列发生器,从土地储备贷款数据中随机抽取一部分作为本研究的样本数据,另外一部分将被用来测试。这种在数据挖掘的数据采集过程中采取随机取样的方式从原始数据中进行数据取样的方法被认为是一种比较简单有效的方法，有效避免样本有偏，即无偏性。

数据采集结果

通过多次使用上述数据处理方法,我们从原始的土地储备项目可研报告中抽取出一张反映土地储备贷款需求的二维关系表,称之为土地储备贷款处理数据集。这个集合共有 74 条记录,每条记录由 36 个字段组成,分别是:收购储备面积、财务内部受益率、动态回收周期、项目投资总额估算、申请贷款额度、银行批复额度、涉及拆迁补偿人口、项目规划用途、总收储成本估算等情况.

字段的有关说明如列表 3 一 1 所示:

表 3 一 1

指 标编码	属性标识	属性名称	数据	取值范围

			类型	
X1	Reserve area	收购储备面积	连续型	35000-7630000 平方米
X2	FNPV	财务净现值	连续型	142-51998 万元
X3	FIRR	财务内部受益率	连续型	0.1241-0.8024
X4	Pt	动态回收周期	连续型	1-2
X5	The estimated total project investment	项目投资总额估算	连续型	3000-85000 万元
X6	Apply for loan amount	申请贷款额度	连续型	2000-50000 万元
X7	The bank approved credits	银行批复额度	连续型	1790-50000 万元
X8	Demolition Compensation population	涉及拆迁补偿人口	离散型	0-1500
X9	Project-planning purposes	项目规划用途	离散型	1,2,3,4
X10	Estimation of total storage cost	总收储成本估算	连续型	3000-85000 万元
X11	The first year of transferring land area	第一年出让土地面积	连续型	12000-1700000 平方米
X12	The first year the proportion of land transfer	第一年出让土地面积比例	连续型	0.3 或 0.4
X13	The first year of income	第一年收入	连续型	1500-51000 万元
X14	Second years of transferring	第二年出让土地面积	连续	18000-2000000 平方米

	land area		型	
X15	Second year the proportion of land transfer	第二年出让土地面积比例	连续型	0.7 或 0.6
X16	Second year of income	第二年收入	连续型	3000-80000 万元
X17	the cash outflow	当年现金流出	连续型	3332-80043 万元
X18	The cash inflow	当年现金流入	连续型	0
X19	Net cash flow	当年净现金流量	连续型	（-80000）-（-3000）万元
X20	The discounted cash flow	当年折现净现金流量	连续型	（-80000）-（-3000）万元
X21	The cumulative net cash flow	当年累计净现金流量	连续型	（-80000）-（-3000）万元
X22	The first annual cash flow	第一年现金流出	连续型	149-4325 万元
X23	The first annual cash inflows	第一年现金流入	连续型	1567-50408 万元
X24	The first annual net cash flow	第一年净现金流量	连续型	1363-46248 万元
X25	The first year of discounted cash flow	第一年折现净现金流量	连续型	1239-42044 万元
X26	The first year cumulative net cash flow	第一年累计净现金流量	连续型	（-41237）-（-1071）万元
X27	The second annual cash outflow	第二年现金流出	连续型	0
X28	The second annual cash inflows	第二年现金流入	连续型	3466-75613 万元
X29	Second years of	第二年净现金流量	连	3466-75613 万元

	net cash flow		续型	
X30	Second years of discounted net cash	第二年折现净现金流量	连续型	2865-62490 万元
X31	The second year cumulative net cash flow	第二年累计净现金流量	连续型	142-25590 万元
X32	Land acquisition and development costs 3% Financial internal rate of return	土地收购开发成本＋3% 财务内部收益率	连续型	0.1171-0.7897
X33	Land acquisition and development costs 3% Increase or decrease	土地收购开发成本＋3% 增减幅度	连续型	（ -0.0861 ） -（-0.0222）
X34	Land revenue 3% Financial internal rate of return	土地收入－3% 财务内部收益率	连续型	0.1170-0.7895
X35	Land revenue 3% Increase or decrease	土地收入－3% 增减幅度	连续型	（ -0.0888 ） -（-0.0225）
X36	Expected return	预期收益	连续型	5034-117446 万元

3.2 数据预处理

3.2.1 数据预处理的必要性

一个完整的数据挖掘系统必须包括数据预处理模块。它以发现任务为目标,以领域知识为指导,用全新的“土地储备贷款风险模型”来组织原来的土地储备贷款数据,摈弃一些与挖掘目标不相关的属性,为数据挖掘内核算法提供干净、准确、更有针对性的数据,从而减少挖掘内核的数据处理量,提高挖掘效率,提高知识发现的起点和知识的准确度。

数据预处理是数据挖掘前的数据准备工作,一方面保证挖掘数据的正确性和有效性,另一方面通过对数据格式和内容的调整,使数据更符合挖掘的需要。其目的在于把一些与数据分析、挖掘无关的项清除掉,为了给挖掘算法提供更高质量的数据。

数据预处理的重要性体现在以下三个方面:

1.数据挖掘算法对要处理的数据集合一般都有一定的要求,比如数据的完整性要好、数据的冗余要少、属性之间的相关性要小。然而,实际系统中的数据一般都具有不完整、冗余性和模糊性,很少能直接满足数据挖掘算法的要求。

2.海量的实际数据中无意义的成分很多,严重影响数据挖掘算法的执行效率,而且其中的噪音干扰还会造成挖掘结果的偏差。

3.数据预处理工作量比纯粹的挖掘过程要大得多,前者约占整个数据挖掘过程的60%左右,而后者只占 10%左右。

因此,对不理想的原始数据进行有效的归纳分析和预处理,已经成为数据挖掘系统实现过程中必须面对的问题。

本研究中,土地储备贷款处理数据集中的数据必须转换成适合数据挖掘算法的形式。但是不同的记录存储方式存在差异,这就需要在数据预处理阶段对它们统一和规范。首先,分析土地储备贷款处理数据集中的数据,非优数据存在以下特点:

1.噪声数据

由于贷款申请表是由贷款申请人和银行信贷人员手工记录的,这就会存在记录错误,它们多由笔误造成。而数据库录入人员为非银行信贷专业人员,对于一些属性值的明显出入不能及时发现,同时录入时偶尔也会将原本正确的数据输错,使得数据具有噪声。带噪声的数据如果不处理则会影响知识发现的准确性。

2.记录形式不统一

不同的记录是由不同的银行信贷人员撰写的,记录形式会有不同,包括记录用词,贷款人情况的描述,描述的简略等方面。

3.大量缺省值

由于土地储备项目可研报告大多是以文字来表述统计数据，没有准确的数据表格做为模板,所以对记录项的取舍以及调查的详略程度有差异,有些认为没有用或用处不大的数据项,他们就会不作记录或作简要的记录,那么有些统计值就会为空,缺损率会很高。

土地储备贷款数据集中的数据具有数量大、记录形式不够统一、大量缺省值存在、简略、存在易错字和易混字等特点,不利于迅速有效地发现所希望得到的信息,因此在进行数据挖掘之前需要针对上述特点对个人住房贷款处理数据集中的数据进行预处理。

3.2.2 数据预处理的内容和方法

数据预处理主要包括:数据清洗、数据集成,数据转换及数据约简。数据清洗是指处理数据中的遗漏数据和脏数据,包括填补遗漏的数据、清除数据中的噪声、剔除异常值以及纠正不一致数据等。数据集成就是将来自多个数据源的异构数据整合到一个完整的数据集.数据转换就是将数据转换成适合数据挖掘的形式,通过寻找数据的特征表示,用维变化方式减少有效变量的数目或找到数据的不变式。数据约简是指通过聚类或删除冗余特征来消除多余数据,从原有大数据集中获得一个精简且完整的数据子集,节省挖掘时间和空间。

这里将根据土地储备贷款数据信息的实际特征,对采集到的处理数据集中的原始数据使用适合的方法进行预处理,生成适合挖掘的目标数据,使其满足下一步数据挖掘工作的需要。

1.数据清洗

1）遗漏数据处理本文采取填充遗漏数据的方法主要有:

①手工填补。当原始资料能够提供,并且数据集不是很大时,此方法可行。本文所获得的个人住房贷款处理数据集的 3922 记录中,“出让计划”和“现金流量”属性中,只有 5 条为空缺值,并且空缺值还可以通过“预期收益”和相邻年份的相关字段的数值计

算或者查找统计年鉴填充,因此这一遗漏数据的填充就采用手工填写的方式。
同类别平均值填补。首先按一定的属性对数据集进行分类,或将具有相同特征的数据聚集起来,然后计算有遗漏的记录所属的类中所有该属性值的默认值、平均值或者同类别平均值,用来替代遗漏值。土地储备贷款处理数据集中,“敏感性分析表”有 4 条有空缺值，是“增减幅度”字段不空的情况,同时有很多属于同“增减幅度”的记录,对这种情况按照“增减幅度”聚集,给予平均值。当所计算出的平均值是噪声数据时,采用线性回归的方法去除噪声。

②噪声数据处理

噪声是指被测变量的随机错误或偏差,包括错误的值或偏离期望的孤立点。本文采用以下两种方法处理噪声数据:
回归方法。利用拟合函数来平滑数据,帮助除去噪声。例如二元线性回归、多元线性回归等。二元线性回归涉及两个变量,可以找出适合两个变童的“最佳”直线,使得一个变量能够预测另一个。多元线性回归是二元线性回归的扩展,它涉及两个以上的变量,使得变量之间存在“最佳”的多维面,在这个面上,能够利用其他变量进行另一个变量的预测。使用回归方法,找出适合的数学方程式,能够帮助消除噪声。
本文的研究中,对于“预期收益”属性,申请贷款时和银行的审批过程都将考虑“第一年出让收入”和“第二年出让收入”这两个主要条件。因此,三者存在一定的关系。可以利用多元线性回归处理噪声数据。在显著性水平$\alpha = 0.05$的情况下，回归方程和回归系数均为显著的,说明$X13, X16$整体上对$X36$有显著的影响。通过式得到的噪声数据的预测值,找到 3922 条记录中有 28 个噪声数据,从而消除这些噪声数据。

人机结合检查方法。这种方法比单纯的人工检查要快。首先由计算机识别并输出那些差异程度大于某个阂值的数据,然后人工审核这些数据,确定孤立点。对于“现金流出”属性,有是负数的记录,共有 16 条,手工改写噪声数据。

③不一致数据处理

数据的不一致包括字段类型的不一致、字段长度的不一致、实体处理(例如各部分对相同的字段采用不同的输入方式)的不一致等,可以通过数据与外部的关联手工处理,比如与原稿校对,或者采用软件工具来发现违反约束条件的数据。
本研究中,“项目规划用途”属性的类型有些是数值型,有些是布尔型,处理成一致的类型是必要的,这里均处理成数值型,用 1 表示商业、住宅用地，2 表示商业用地，3 表示住宅用地，4 表示综合用地。“拆迁人口”属性有些是布尔型,有些是数值型,这里均处理成布尔型,用 0 表示没有,1 表示有。

2. 数据集成

本文数据来源及数据本身的特点决定了本研究主要通过处理数据的冗余解决数据集成问题,而数据冗余的处理体现在处理记录冗余和属性冗余。

①记录冗余的处理

对于同一数据集,存在两条或多条相同的记录,就是记录冗余。在土地储备贷款处理数据集中,存在同一贷款申请土地在同一时期或不同时期有多土地贷款的情况,而本研究数据挖掘目的是对土地储备风险状况进行评估,这样就会出现记录冗余。对此,应使用相应的合并记录函数,合并贷款记录,结果共删除 4 条冗余记录.

②属性冗余的处理

若一个属性可以从其他属性推演出来,那么它就是冗余属性。数据集成往往导致数据冗余,如同一属性多次出现、同一属性命名不一致等。对于属性冗余,本研究采用以下

3 种处理方法:

逻辑上直观判断。在进行数据预处理时有些属性可以根据逻辑上直观的判断决定取舍,如第一年出让面积、第二年出让面积、预计收储面积。对于直观上不好判断的情况,可以利用数理统计中的相关性分析方法检测数值属性是否相关。

χ^2 方法考察属性间的相关性。可以用统计分布的扩方法来考察属性间的相关性,除去相关性高的属性。本研究中数据预处理的数据相关性分析也用到
了这个方法。

χ^2 统计检验是对提供样本的频数分布的总体分布是否服从某种理论分布或某种假设分布所作的假设检验。

在土地储备贷款处理数据集中,经过计算得到的 P 值接近 1,认为这两个属性高度相关的,因此除去，是冗余属性,将其删除。

信息增益考察属性间的相关性。除上述两种方法外,本研究还通过前面已经介绍过的信息增益考察属性间的相关性,消除信息增益小的属性。我们计算定义 S 中样本的每个属性的信息增益,设用于识别弱相关性的属性相关阈值为 aD,若属性的信息增益小于该阈值则被认为是弱相关的,应删除。例如，土地规划用途中的工业用地以及商住综合用地，这两个属性由于信息量太小,而无法对决策产生影响。

列去噪，发现一种情况。

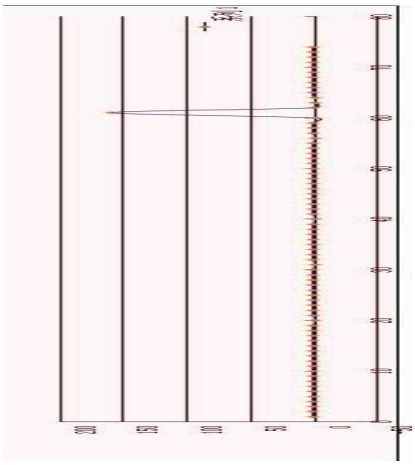


图 3-1：第一年与第二年土地出让面积是收储总面积的 92%-108%的样本只有 42 个

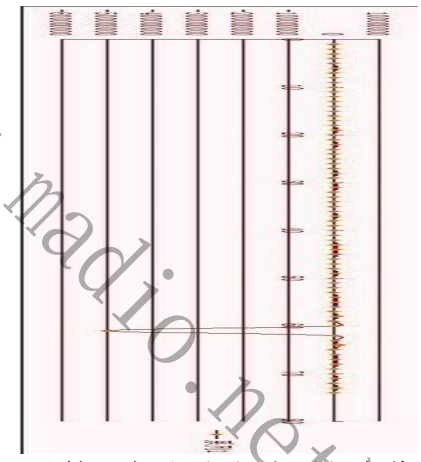


图 3-2：第一年出让土地面积占收储总面积的比例

序号	第一年出让土地+第二年出让土地	收购面积-出让总面积	第一年出让总面积/总收储
2	122797.2	-2000	0.316556675
1	300000	0	0.3
3	239765.92	0	0.400000008
7	192000	-157000	0.342857143
4	75333.8	0	0.299998142
10	995390	1405810	0.165815426
9	306642.36	192192.64	0.245886796
8	549778.91	0	0.399999993
11	136334.02	0	0.400000015
13	80000	0	0.4
14	191847.43	0	0.39999999
5	535300	0	0.4
6	506544	0	0.4
12	360000	90000	0.32
24	157637	0	0.4
36	179755	0	0.4
30	600000	0	0.4
26	188446	0	0.4
16	713914.9	305963.53	0.28
27	30668.2	276000	0.0400018
18	249536	0	0.400002404
35	463095.6	308730.4	0.24
32	579300	0	0.4
38	310075	80	0.399896826
31	500000	130200	0.317359568
28	850975.93	0	0.399999998
25	102400.79	0	0.400000039
15	684342	306383.68	0.276299288
21	450000	0	0.4
23	550000	350000	0.244444444
37	400200	2610638	0.053167922
33	845670.9	200518.6	0.323333736
39	244822.35	-121.13	0.400198005
41	135334.02	49026.9	0.295798101
29	570750	190250	0.3
17	952730.87	0	0.400000002
34	611000	329000	0.26
40	498916	125083.38	0.319818266
19	800400	0	0.4
42	192000	108000	0.04
20	691000	0	0.4
45	311200	0	0.4
22	1554007.77	666003.33	0.280000001
44	660000	540000	0.22
43	735761	638259	0.214192224
49	1500000	750	0.3998001
54	278021	0.39	0.399999439
62	310075	80	0.399896826
51	2680909.4	670227.36	0.319999999
65	940000	0	0.4
46	407203	47500	0.358214483
52	556402.78	0	0.399999996
63	2200000	0	0.4
55	495228.18	504771.82	0.19809127
56	495228.18	504771.82	0.19809127
58	700000	1358362	0.136030494
47	3185000	0	0.4
50	2907955.62	4722044.38	0.152448526
53	336000	144000	0.28
60	650000	1618011.34	0.114637875
61	503987721.3	-502728381.7	160.0799994
59	758790.46	531229.32	0.235280253
64	928000	1877347.36	0.13231873
57	4244948.53	12410821.47	0.101945417
67	382846	0	0.4
66	1334000	2008337	0.159648773
68	660000	340000	0.264
69	1465610	523540	0.294720861
48	2204677.69	0	0.400000002
70	422668.78	105627.19	0.398768365
73	720000	480006	0.2399988
71	406369.75	9312.33	0.391038988
74	1158430.4	1228659.6	0.194115915
72	420283.37	74630.63	0.33967655

综上所述，去掉了 X11、X12、X13 和 X14 等列。

对原始表格 36 列做相关性检验并给出残差检验。保留相关性低于 90%的列，得到以下 X1、X2、X3、X4、X8、X9、X11、X12、X14、X17、X22、X31、X35、X36 共 14 列。

表 3-3

相关列	相关系数
<e,f>	-0.98136
<e,ah>	0.991277
<e,aj>	0.991148
<f,ah>	-0.964935
<f,aj>	-0.964617
<g,h>	0.985104
<g,l>	1
<g,o>	0.988463
<g,r>	0.98867
<g,u>	-0.999946
<g,v>	-0.999946
<g,w>	-0.999946
<g,y>	0.98849
<g,z>	0.985923
<g,aa>	0.985922
<g,ab>	-0.986929
<g,ad>	0.9887
<g,ae>	0.988636
<g,af>	0.988636
<g,al>	0.988648
<h,i>	0.959918
<h,l>	0.985104
<h,o>	0.963422
<h,r>	0.963583
<h,u>	-0.98399
<h,v>	-0.98399
<h,w>	-0.98399
<h,y>	0.963462
<h,z>	0.958109
<h,aa>	0.958109
<h,ab>	-0.983028
<h,ad>	0.963627
<h,ae>	0.963562
<h,af>	0.963562
<h,al>	0.963578
<l,o>	0.988463
<l,r>	0.98867
<l,u>	-0.999946
<l,v>	-0.999946
<l,w>	-0.999946
<l,y>	0.98849
<ah,aj>	0.999996

相关列	相关系数
<l,aa>	0.985922
<l,ab>	-0.986929
<l,ad>	0.9887
<l,ae>	0.988636
<l,af>	0.988636
<l,al>	0.988648
<n,q>	-1
<o,r>	0.999744
<o,u>	-0.988554
<o,v>	-0.988554
<o,w>	-0.988554
<o,y>	0.999993
<o,z>	0.999718
<o,aa>	0.999718
<o,ab>	-0.951146
<o,ad>	0.99974
<o,ae>	0.999724
<o,af>	0.999724
<o,al>	0.999908
<r,u>	-0.988761
<r,v>	-0.988761
<r,w>	-0.988761
<r,y>	0.999734
<r,z>	0.999419
<r,aa>	0.999418
<r,ab>	-0.951845
<r,ad>	0.999993
<r,ae>	0.99998
<r,af>	0.99998
<r,al>	0.999959
<s,x>	0.983933
<u,v>	1
<u,w>	1
<u,y>	-0.988573
<u,z>	-0.986214
<u,aa>	-0.986214
<u,ab>	0.98675
<u,ad>	-0.988784
<u,ae>	-0.98872
<u,af>	-0.98872
<u,al>	-0.988739
<v,w>	1

相关列	相关系数
<v,y>	-0.988573
<v,z>	-0.986214
<v,aa>	-0.986214
<v,ab>	0.98675
<v,ad>	-0.988784
<v,ae>	-0.98872
<v,af>	-0.98872
<v,al>	-0.988739
<w,y>	-0.988573
<w,z>	-0.986214
<w,aa>	-0.986214
<w,ab>	0.98675
<w,ad>	-0.988784
<w,ae>	-0.98872
<w,af>	-0.98872
<w,al>	-0.988739
<y,z>	0.999709
<y,aa>	0.999709
<y,ab>	-0.951192
<y,ad>	0.999744
<y,ae>	0.999728
<y,af>	0.999728
<y,al>	0.9999
<z,aa>	1
<z,ad>	0.999412
<z,ae>	0.999401
<z,af>	0.999401
<z,al>	0.9996
<aa,ad>	0.999412
<aa,ae>	0.999401
<aa,af>	0.999401
<aa,al>	0.9996
<ab,ad>	-0.951895
<ab,ae>	-0.951781
<ab,af>	-0.951781
<ab,al>	-0.951623
<ad,ae>	0.999987
<ad,af>	0.999987
<ad,al>	0.999953
<ae,af>	1
<ae,al>	0.999939
<af,al>	0.999939

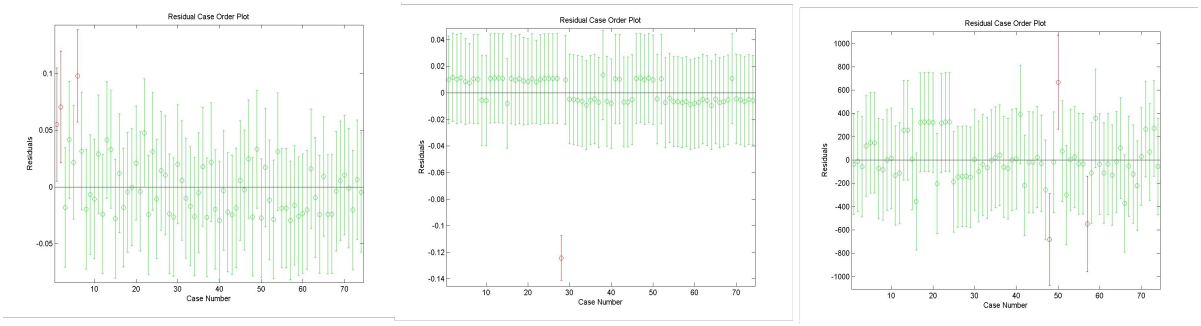


图 3-3

图中红色的数据项为异常值，在拟合优度 90%以上的时候，可以选择删除掉该项。

列联表分析

表 3-4

相关矩阵

		收购储备面积（平方米）	财务净现值（FNPV）	财务内部受益率（FIRR）	动态回收周期（Pt）	涉及拆迁补偿人口（户）	项目规划用途	当年现金流	第一年现金流量	第二年累计净现金流量	土地收入—3%增减幅度	预期收益（万元）
相关	收购储备面积（平方米）	1.000	.266	.115	-.161	-.081	-.116	.191	.150	.370	.387	.276
	财务净现值（FNPV）	.266	1.000	.281	-.342	-.028	-.069	.482	.456	.774	-.087	.660
	财务内部受益率（FIRR）	.115	.281	1.000	-.981	.145	.004	-.043	-.008	.385	-.165	-.043
	动态回收周期（Pt）	-.161	-.342	-.981	1.000	-.125	.009	-.004	-.030	-.454	.156	-.014
	涉及拆迁补偿人口（户）	-.081	-.028	.145	-.125	1.000	.074	.025	.048	.011	.055	-.018
	项目规划用途	-.116	-.069	.004	.009	.074	1.000	-.129	-.162	-.045	-.203	-.059
	当年现金流	.191	.482	-.043	-.004	.025	-.129	1.000	.984	.604	.033	.687
	第一年现金流量	.150	.456	-.008	-.030	.048	-.162	.984	1.000	.574	-.003	.651
	第二年累计净现金流量	.370	.774	.385	-.454	.011	-.045	.604	.574	1.000	-.082	.851

	土地收入－3%增减幅度	.387	-.087	-.165	.156	.055	-.203	.033	-.003	-.082	1.000	.041
	预期收益(万元)	.276	.660	-.043	-.014	-.018	-.059	.687	.651	.851	.041	1.000
Sig.单侧	收购储备面积（平方米）		.011	.165	.085	.246	.162	.051	.100	.001	.000	.009
	财务净现值（FNPV）	.011		.008	.001	.406	.279	.000	.000	.000	.230	.000
	财务内部受益率（FIRR）	.165	.008		.000	.109	.486	.358	472	.000	.080	.357
	动态回收周期（Pt）	.085	.001	.000		.143	.468	.486	.399	.000	.092	.454
	涉及拆迁补偿人口（户）	.246	.406	.109	.143		.267	.417	.344	.462	.320	.441
	项目规划用途	.162	.279	.486	.468	.267		.136	.084	.351	.041	.310
	现金流量表	.051	.000	.358	.486	.417	.136		.000	.000	.389	.000
	V10	.100	.000	.472	.399	.344	.084	.000		.000	.491	.000
	V11	.001	.000	.000	.000	.462	.351	.000	.000		.244	.000
	V12	.000	.230	.080	.092	.320	.041	.389	.491	.244		.365
	预期收益(万元)	.009	.000	.357	.454	.441	.310	.000	.000	.000	.365	

继续对列去噪，得到 X1、X2、X3、X4、X8、X9、X17、X22、X31、X35、X36 共 11 列。通过关联系数以及显著性分析得到最终的属性列为：X1、X2、X3、X8、X9、X17、X35、X36 共 8 列。

3.数据转换

本研究使用了 3 种数据转换技术:

(1)数据规格化

数据转换主要是数据规格化。规格化是属性值量纲的归一化处理,目的是消除数值型属性因大小不一而造成挖掘结果的偏差.

(2)数据类型转换

数据的处理中,比起字符型数据,数值型数据占有更小的存储空间,具有更快的计算速度,因此,在数据挖掘的数据形成过程中经常将字符型数据转换成数值型数据。在个人住房贷款处理数据集中,将“商业、住宅用地“、“商业用地”、“住宅用地”、“综合用地”的字符型转换成数值型存储，分别转换为:1、2、3、4.

(3)属性构造

属性构造是由给定的属性构造和添加新的属性。一是根据国际上将“土地收入-3%的增减幅度”作为控制土地储备贷款风险的衡量办法之一,这样就能准确的反映土地储

备贷款状况。也因此“土地储备成本+3%的财务内部收益率”、“土地储备成本+3%的增减幅度”和“土地收入-3%的财务内部收益率”都成为冗余属性,将它们从土地储备贷款处理数据集中删除。

4.数据约简

常用的数据约简方法有:属性约简、数据块约简、离散化与概念分层。属性约简是指通过删除跟挖掘任务无关的或冗余的属性(或维)来减少数据规模。数据块约简是指通过选择较小的数据表示形式来替代原数据以减少数据量.在数据采集过程中,本文选择并构造了与挖掘目的有一定关系的属性,且通过属性冗余的处理,删除了与挖掘任务不相关或弱相关的属性,已经基本实现了土地储备贷款处理数据集属性及数据块约简的目的。这里主要进行离散化与概念分层的操作。连续属性离散化就是在特定的连续属性的值域范围内设定若干个离散化划分点,将属性的值域范围划分成一些离散化区间,再用不同的符号或整数值代表属于每个区间的属性值。概念分层是用较高层的概念替换原始数据或较低层的概念。连续属性的离散化和概念分层大大浓缩了数据库记录。尽管这种泛化使得细节丢失,但泛化后的数据更有意义并容易理解,有助于挖掘不同抽象层次的模式知识。

土地储备数据集的预处理结果：

表 3-5

指标编码	属性标识	属性名称	数据类型	预处理方法及结果
X1	Reserve area	收购储备面积	离散型	将收购储备面积按照大小分为三类 收购储备面积低 = 1 收购储备面积中 = 2 收购储备面积高 = 3
X2	FNPV	财务净现值	离散型	将财务净现值按照大小分为三类 财务净现值低 = 1 财务净现值中 = 2 财务净现值高 = 3
X3	FIRR	财务内部受益率	离散型	将财务内部收益率按照值大小分为三类 财务内部收益率低 = 1 财务内部收益率中 = 2 财务内部收益率高 = 3
X8	Demolition Compensation population	涉及拆迁补偿人口	离散型	将涉及到的拆迁补偿人口分为有拆迁人口和无拆迁人口两类 无拆迁人口 = 0

				有拆迁人口 = 1
X9	Project planning purposes	项目规划用途	离散型	将项目规划用途较少的属性样本归类到其它类似属性中，并且将其删除 商业、住宅用地=1 商业用地=2 住宅用地=3 综合用地=4
X17	the cash outflow	当年现金流出	离散型	将当年现金流出按照大小分为三类 当年现金流出低 = 1 当年现金流出中 = 2 当年现金流出高 = 3
X35	Land revenue 3% Increase or decrease	土地收入－3%增减幅度	离散型	将该属性按照数值大小分为高中低三类 增减幅度低 = 1 增减幅度中 = 2 增减幅度高 = 3
X36	Expected return	预期收益	离散型	将预期收益按照大小分为三类 预期收益低 = 1 预期收益中 = 2 预期收益高 = 3

对属性列为：X1、X2、X3、X8、X9、X17、X35、X36 共 8 列，进行因子分析和提取主成分的工作。分别得到相关矩阵、碎石图、成分矩阵、成分图等。接着对 8 个属性值进行 K—均值聚类。得到初始聚类中心、ANOVA、每个聚类中的案例数。

表 3-6

相关矩阵

		收购储备面积（平方米）	财务净现值（FNPV）	财务内部受益率（FIRR）	涉及拆迁补偿人口（户）	项目规划用途	当年现金流出	土地收入－3%增减幅度	预期收益（万元）
相关	收购储备面积（平方米）	1.000	.266	.115	-.081	-.116	.191	.387	.276
	财务净现值（FNPV）	.266	1.000	.281	-.028	-.069	.482	-.087	.660
	财务内部受益率（FIRR）	.115	.281	1.000	.145	.004	-.043	-.165	-.043

	涉及拆迁补偿人口（户）	-.081	-.028	.145	1.000	.074	.025	.055	-.018
	项目规划用途	-.116	-.069	.004	.074	1.000	-.129	-.203	-.059
	当年现金流出	.191	.482	-.043	.025	-.129	1.000	.033	.687
	土地收入－3％增减幅度	.387	-.087	-.165	.055	-.203	.033	1.000	.041
	预期收益（万元）	.276	.660	-.043	-.018	-.059	.687	.041	1.000
Sig.（单侧）	收购储备面积（平方米）		.011	.165	.246	.162	.051	.000	.009
	财务净现值（FNPV）	.011		.008	.406	.279	.000	.230	.000
	财务内部受益率（FIRR）	.165	.008		.109	.486	.358	.080	.357
	涉及拆迁补偿人口（户）	.246	.406	.109		.267	.417	.320	.441
	项目规划用途	.162	.279	.486	.267		.136	.041	.310
	当年现金流出	.051	.000	.358	.417	.136		.389	.000
	土地收入－3％增减幅度	.000	.230	.080	.320	.041	.389		.365
	预期收益（万元）	.009	.000	.357	.441	.310	.000	.365	

表 3-7

公因子方差		
	初始	提取
收购储备面积(平方米)	1.000	.655
财务净现值（FNPV）	1.000	.785
财务内部受益率（FIRR）	1.000	.904

涉及拆迁补偿人口（户）	1.000	.916
项目规划用途	1.000	.338
当年现金流出	1.000	.753
土地收入－3%增减幅度	1.000	.803
预期收益（万元）	1.000	.857
提取方法：主成份分析。		

表 3-8

解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	2.404	30.053	30.053	2.404	30.053	30.053
2	1.438	17.970	48.022	1.438	17.970	48.022
3	1.162	14.527	62.549	1.162	14.527	62.549
4	1.006	12.581	75.130	1.006	12.581	75.130
5	.904	11.295	86.425			
6	.447	5.583	92.008			
7	.414	5.173	97.181			
8	.226	2.819	100.000			
提取方法：主成份分析。						

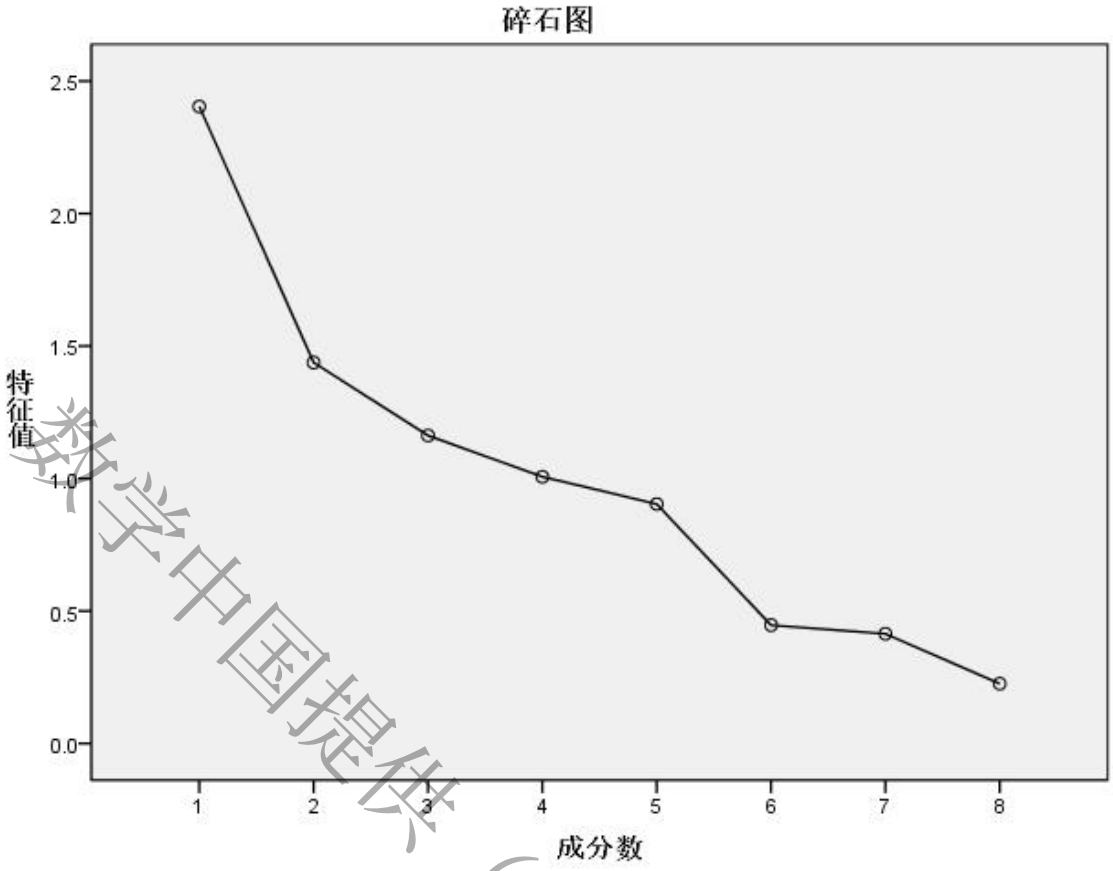
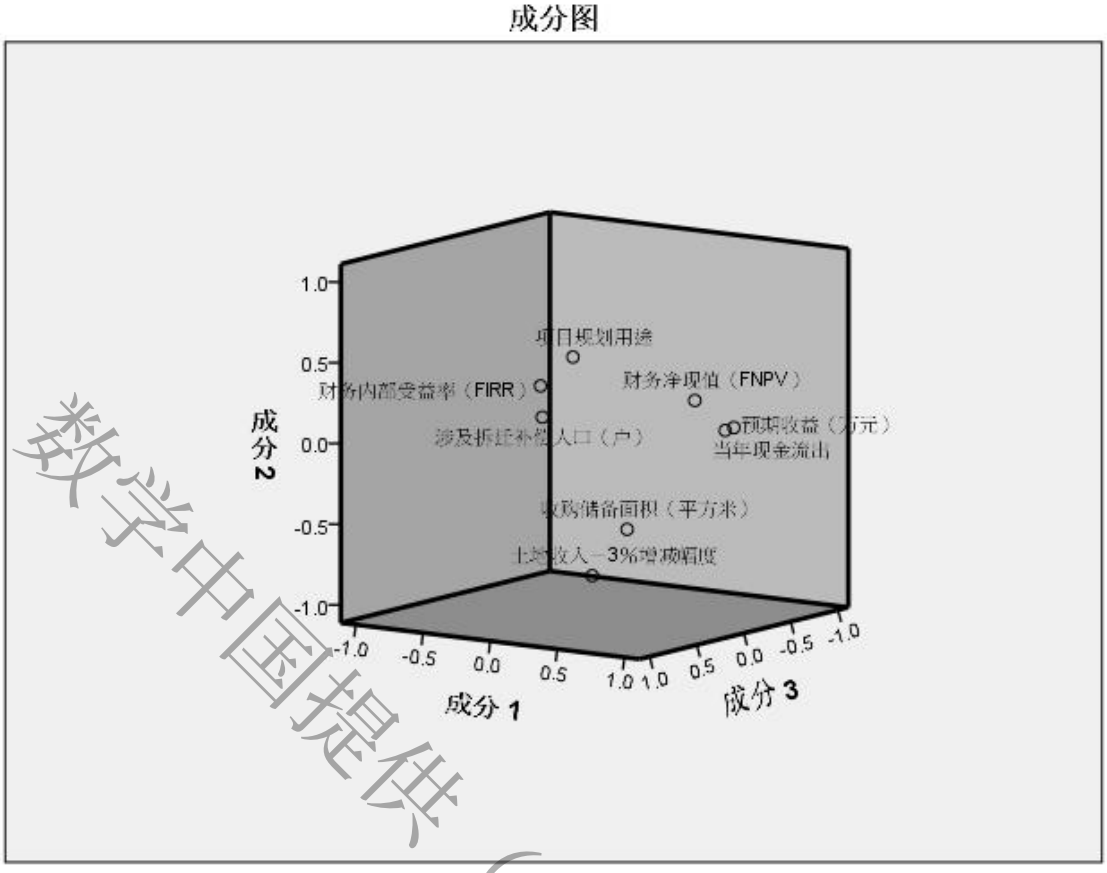


图 3-4

表 3-9

成份矩阵[a]				
	成份			
	1	2	3	4
收购储备面积(平方米)	.506	-.475	.377	-.177
财务净现值（FNPV）	.810	.314	.091	-.150
财务内部受益率（FIRR）	.130	.432	.761	-.348
涉及拆迁补偿人口（户）	-.034	.184	.507	.790
项目规划用途	-.215	.455	-.076	.281
当年现金流出	.796	.077	-.252	.225
土地收入－3%增减幅度	.148	-.818	.237	.237
预期收益（万元）	.879	.107	-.225	.148
提取方法 ：主成份。				
a. 已提取了 4 个成份。				



无

图 3-5

表 3-10

成份得分系数矩阵

	成份			
	1	2	3	4
收购储备面积(平方米)	.210	-.330	.324	-.176
财务净现值（FNPV）	.337	.218	.078	-.149
财务内部受益率（FIRR）	.054	.300	.655	-.345
涉及拆迁补偿人口（户）	-.014	.128	.436	.785
项目规划用途	-.089	.317	-.065	.279
当年现金流出	.331	.053	-.217	.224
土地收入－3%增减幅度	.062	-.569	.204	.235
预期收益（万元）	.366	.074	-.193	.147

提取方法 ：主成份。
构成得分。

表 3-11

成份得分协方差矩阵				
成份	1	2	3	4
1	1.000	.000	.000	.000
2	.000	1.000	.000	.000
3	.000	.000	1.000	.000
4	.000	.000	.000	1.000
提取方法 ：主成份。 构成得分。				

表 3-12

初始聚类中心				
	聚类			
	1	2	3	4
收购储备面积(平方米)	16655770.00	7630000.00	3342337.00	35000.00
财务净现值（FNPV）	16875.12	12501.84	21028.85	963.95
财务内部受益率（FIRR）	.35	.36	.36	.18
涉及拆迁补偿人口（户）	.00	.00	1320.00	.00
当年现金流出	44448.31	31132.00	54144.70	7999.60
土地收入－3%增减幅度	.03	-.03	-.03	-.02
项目规划用途	1.00	1.00	1.00	1.00
预期收益（万元）	73177.26	53165.14	90000.00	10800.00

表 3-15

最终聚类中心				
	聚类			
	1	2	3	4
收购储备面积(平方米)	16655770.00	7630000.00	2506002.04	584961.69
财务净现值（FNPV）	16875.12	12501.84	12913.93	7387.33
财务内部受益率（FIRR）	.35	.36	.34	.30

涉及拆迁补偿人口（户）	.00	.00	120.00	158.66
当年现金流出	44448.31	31132.00	38962.62	23792.28
土地收入－3%增减幅度	.03	-.03	-.03	-.03
项目规划用途	1.00	1.00	2.17	1.83
预期收益（万元）	73177.26	53165.14	62365.42	36696.82

表 3-16

ANOVA

	聚类		误差		F	Sig.
	均方	df	均方	df		
收购储备面积(平方米)	107527941765328.640	3	156097665440.193	68	688.850	.000
财务净现值（FNPV）	130582563.270	3	57294158.761	68	2.279	.087
财务内部受益率（FIRR）	.007	3	.017	68	.446	.721
涉及拆迁补偿人口（户）	19932.725	3	109720.163	68	.182	.908
当年现金流出	876494599.605	3	360680752.913	68	2.430	.073
土地收入－3%增减幅度	.001	3	.000	68	10.439	.000
项目规划用途	.890	3	1.440	68	.618	.606
预期收益（万元）	2566273857.989	3	781460936.645	68	3.284	.026

F 检验应仅用于描述性目的，因为选中的聚类将被用来最大化不同聚类中的案例间的差别。观测到的显著性水平并未据此进行更正，因此无法将其解释为是对聚类均值相等这一假设的检验。

表 3-17

每个聚类中的案例数		
聚类	1	1.000
	2	1.000
	3	12.000
	4	58.000
有效		72.000

缺失	84.000

得到很好的分类效果。

接下来通过回归分析和残差图对 74 个样本去噪，发现样本 5、6、22、24、26、36、38、45、56、71、72、74 可能对决策树分类的准确性有影响。

土地储备风评估模型分析：

表 3-18 各属性分类后的信息熵

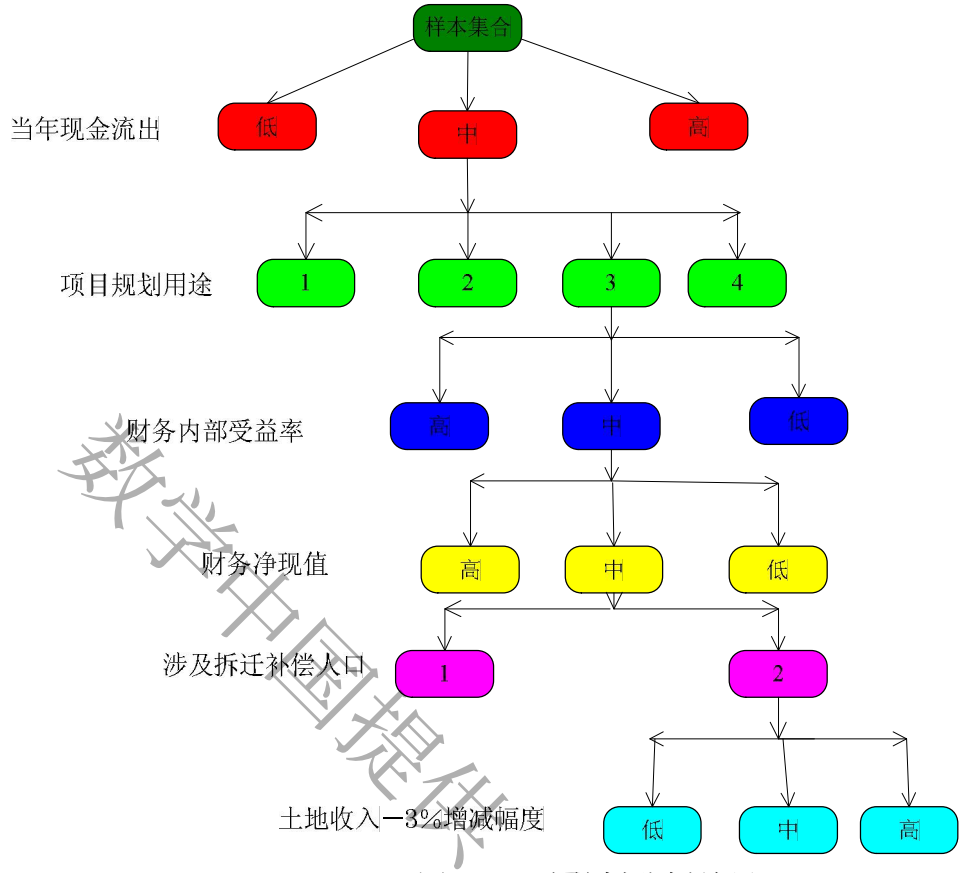
属性	信息熵
收购储备面积	0.954434002924965
财务净现值	1.293697716489340
财务内部收益率	0.616748259826391
涉及拆迁补偿人口	0.870864469235364
当年现金流出	1.316384182158285
土地收入-3%增减幅	0.183122068301373
预期收益	1.298794940695399
项目规划用途	1.128492401573571

从上表中可以看出当年现金流出的信息熵值最大，根据前面章节综述，本文选取当年现金流出信息作为基准，继续对其他属性做条件熵。条件熵及信息增益如下表所示。

表 3-19 条件熵及信息增益

	X1	X2	X3	X8	X35	X36	X9
条件熵	0.5907	3.9951	6.2429	3.3171	2.8131	2.0481	6.3014
Gain	0.7257	2.6788	4.9265	2.0007	1.4967	0.7317	4.9850
ratio	0.0414	0.1527	0.2808	0.1140	0.0853	0.0417	0.2841
weight	4	15	28	11	9	4	28

从信息增益率中可以看出收储面积与预期收益的信息增益率非常小，信息增益率的现实含义是从每个属性获得的信息量,代表该属性对结果类划分贡献的大小,而信息增益率太小就意味着对结果集的分类影响非常小。因此,为了避免决策树过于庞大,删除信息增益率太小的这两个属性,只对其他属性进行分析，构造决策树。



图：3-6 决策树分析结果

决策树的分析结果表明当年现金流出是决策树分枝的最重要因素，次为项目规划用途。它们都是影响土储方案风险评估的重要依据。

由于题中所给的数据都没有分类这些土地资源储备案例的盈利情况，所以在风险评估时，虽然知道考虑的属性的优先级，但不能明确知道利好的情况。只能通过聚类分析计算出各个项目的风险大小，下表为由配合决策树统计出的十个风险最大项目。

1	2	23	47	48	49	54	66	72	74

观察这些项目，不难发现从 54 往后的项目当年现金流出的数值较大，48、49 两项数据的收益率较小，且 49 中的收储面积很大，但财务净现值较小体现出 49 有高风险。1 和 66 受土地收入—3%增减幅度影响较大，且 66 当年的现金流出较大，1 的财务内部收益率较小。而且表中部分项目都有一个共性就是部分项目涉及要补贴拆迁人口。从中可以得出决策树可以有效引导土储相关部门去控制风险。同样我们用 MATLAB 和 SPSS 进行了相同的决策树分析得到了相近的结果，用决策树分析结果图展示出来。

四、模型二的建立求解与结论分析

4. 基于 Logistic 回归的土地储备方案的评级模型

4.1.1 属性的编码与建模变量的生成

训练样本集合包含 8 个特征属性用于建立 Logistic 回归模型.

对于类别属性 Y，由于只分为 2 个类别，因此可对其编码如下：

$$Y=\begin{cases}1, \text{风险小}\\ 0, \text{风险大}\end{cases}:$$

经过上述处理后，8 个特征属性 X1、X2、X3、X8、X9、X17、X35、X36 作为建立 Logistic 回归模型的自变量，属性 Y 作为模型的因变量。

4. 1. 2 模型的建立与结果

我们以 8 个变量作为自变量，以 Y 作为因变量建立 Logistic 回归方程。在这里我们针对一个土地储备贷款是“风险小”的概率 p(Y =1)建立模型如下:

$$\ln\frac{p(Y=1|X)}{1-P(Y=1|X)}=\beta_0+\beta^T X$$

由于用于建模的变量比较多，难免有些自变量可能对因变量的解释没有贡献，并且某些自变量之间可能存在较强的线性关系，因此在建立 Logistic 回归模型时我们采取“向后逐步选择(Backwark: Wald)”方法，即根据 Wald 统计量的概率进行提出变量的检验。

我们利用 SPSS 软件进行建模，模型参数的估计结果如下表:

表 4-1

模型系数的综合检验

		卡方	df	Sig.
步骤 1	步骤	17.853	10	.057
	块	17.853	10	.057
	模型	17.853	10	.057
步骤 2[a]	步骤	-.099	1	.753
	块	17.754	9	.038
	模型	17.754	9	.038
步骤 3[a]	步骤	-1.409	1	.235
	块	16.345	8	.038
	模型	16.345	8	.038
步骤 4[a]	步骤	-.495	1	.482
	块	15.850	7	.027
	模型	15.850	7	.027
步骤 5[a]	步骤	-.858	1	.354
	块	14.992	6	.020
	模型	14.992	6	.020
步骤 6[a]	步骤	-1.047	1	.306
	块	13.946	5	.016
	模型	13.946	5	.016

步骤 7[a]	步骤	-7.962	3	.047
	块	5.984	2	.050
	模型	5.984	2	.050
步骤 8[a]	步骤	-1.672	1	.196
	块	4.312	1	.038
	模型	4.312	1	.038

a. 负卡方值表示卡方值已从上一步中减小。

表 4-2

模型汇总

步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	40.170[a]	.220	.397
2	40.269[a]	.219	.395
3	41.679[a]	.203	.367
4	42.173[a]	.198	.357
5	43.031[a]	.188	.340
6	44.078[a]	.176	.318
7	52.040[b]	.080	.144
8	53.712[b]	.058	.105

a. 因为已达到最大迭代次数，所以估计在迭代次数 20 处终止。 无法找到最终解。

b. 因为参数估计的更改范围小于 .001，所以估计在迭代次数 5 处终止。

表 4-3

= Hosmer 和 Lemeshow 检验 =

步骤	卡方	df	Sig.
1	4.740	8	.785
2	4.878	8	.771
3	10.198	8	.251
4	12.003	8	.151
5	3.516	8	.898
6	8.737	8	.365
7	11.620	8	.169
8	8.366	8	.399

表 4-4

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)
步骤1[a]	收购储备面积(平方米)	.000	.000	1.369	1	.242	1.000
	财务净现值(FNPV)	.000	.000	.030	1	.863	1.000
	财务内部受益率(FIRR)	10.134	7.004	2.094	1	.148	25184.478
	涉及拆迁补偿人口(户)	-.003	.002	3.568	1	.059	.997
	项目规划用途			5.485	3	.140	
	项目规划用途(1)	2.316	1.003	5.338	1	.021	10.140
	项目规划用途(2)	2.091	1.766	1.402	1	.236	8.096
	项目规划用途(3)	22.879	9812.353	.000	1	.998	8634508456.772
	当年现金流出	.000	.001	.544	1	.461	1.000
	土地收入－3%增减幅度	-118.144	84.111	1.973	1	.160	.000
	预期收益（万元）	.000	.000	.515	1	.473	1.000
	常量	-4.304	3.621	1.413	1	.235	.014
步骤2[a]	收购储备面积（平方米）	.000	.000	1.385	1	.239	1.000
	财务内部受益率（FIRR）	10.134	7.038	2.073	1	.150	25178.100
	涉及拆迁补偿人口（户）	-.003	.002	3.616	1	.057	.997
	项目规划用途			5.575	3	.134	
	项目规划用途(1)	2.335	1.002	5.431	1	.020	10.326
	项目规划用途(2)	2.081	1.762	1.396	1	.237	8.016
	项目规划用途(3)	22.868	9815.398	.000	1	.998	8537823352.367
	当年现金流出	.000	.000	1.577	1	.209	1.000
	土地收入－3%增减幅度	-119.002	84.417	1.987	1	.159	.000
	预期收益（万元）	.000	.000	1.971	1	.160	1.000
	常量	-4.317	3.638	1.408	1	.235	.013
步骤3[a]	财务内部受益率（FIRR）	8.438	6.676	1.598	1	.206	4620.842
	涉及拆迁补偿人口（户）	-.003	.002	3.335	1	.068	.997
	项目规划用途			4.974	3	.174	

	项目规划用途(1)	2.024	.924	4.800	1	.028	7.567
	项目规划用途(2)	1.880	1.692	1.234	1	.267	6.551
	项目规划用途(3)	22.546	10375.861	.000	1	.998	6185725408.095
	当年现金流出	.000	.000	.772	1	.380	1.000
	土地收入－3%增减幅度	-29.475	38.968	.572	1	.449	.000
	预期收益（万元）	.000	.000	1.056	1	.304	1.000
	常量	-1.299	2.268	.328	1	.567	.273
步骤4[a]	财务内部受益率（FIRR）	8.442	6.634	1.619	1	.203	4636.491
	涉及拆迁补偿人口（户）	-.003	.002	3.172	1	.075	.997
	项目规划用途			4.665	3	.198	
	项目规划用途(1)	1.906	.900	4.487	1	.034	6.724
	项目规划用途(2)	1.778	1.653	1.157	1	.282	5.916
	项目规划用途(3)	22.400	10709.798	.000	1	.998	5347846640.431
	当年现金流出	.000	.000	.784	1	.376	1.000
	预期收益（万元）	.000	.000	1.070	1	.301	1.000
	常量	-.495	1.971	.063	1	.802	.610
步骤5[a]	财务内部受益率（FIRR）	3.965	4.009	.978	1	.323	52.698
	涉及拆迁补偿人口（户）	-.003	.001	3.471	1	.062	.997
	项目规划用途			4.482	3	.214	
	项目规划用途(1)	1.855	.888	4.364	1	.037	6.393
	项目规划用途(2)	1.511	1.515	.994	1	.319	4.529
	项目规划用途(3)	21.904	11016.322	.000	1	.998	3256599415.544
	预期收益（万元）	.000	.000	3.637	1	.057	1.000
	常量	.707	1.441	.240	1	.624	2.027
步骤6[a]	涉及拆迁补偿人口（户）	-.002	.001	2.908	1	.088	.998
	项目规划用途			4.319	3	.229	
	项目规划用途(1)	1.842	.888	4.302	1	.038	6.310
	项目规划用途(2)	1.177	1.399	.708	1	.400	3.245
	项目规划用途(3)	21.217	11741.95	.000	1	.999	163802882.112
	预期收益（万元）	.000	.000	4.341	1	.037	1.000

	常量	1.941	.854	5.168	1	.023	6.966
步骤7[a]	涉及拆迁补偿人口（户）	-.001	.001	1.804	1	.179	.999
	预期收益（万元）	.000	.000	4.606	1	.032	1.000
	常量	3.135	.720	18.951	1	.000	22.979
步骤8[a]	预期收益（万元）	.000	.000	4.439	1	.035	1.000
	常量	2.874	.667	18.575	1	.000	17.713

a. 在步骤 1 中输入的变量：收购储备面积（平方米），财务净现值（FNPV），财务内部受益率（FIRR），涉及拆迁补偿人口（户），项目规划用途，当年现金流出，土地收入－3%增减幅度，预期收益（万元）.

从而模型可以写成:

$$\ln\frac{p}{1-p}=0.542+5.452X3-0.001X8-0.558X9-89.375X35$$

其中 P 为一个土地储备是小风险的概率。

根据模型的估计概率的直方图，我们将分类分界点设置为 0. 4，即当 p ＜ 0.4 时，判定

为大风险;当 p ≥ 0.4 时判定为小风险。

从上表可以看出决策树模型的整体表现是相对比较突出的。这进一步表明了我们用决策树方法建立的土地储备风险评级模型具有很高的实用价值。

五、参考文献

[1] 于卓.应用决策树构建个人住房贷款风险评估模型[D].东北财经大学,2007.
[2] 巩吉璋.决策树分类算法在银行个人信用评级中的应用[D].暨南大学,2008.
[3] 邓舒放.基于决策树_神经网络的个人信用评估组合模型的构建[D].湖南大学,2012.