

课程设计报告

**美国金县二手房价的分析与预测**

**姓 名：**计科1905谷平阔

**学 号：**19851226

**专 业：**计算机科学与技术

**系 别：**计算机与信息技术学院

**指导教师：**王志海

**2022年6月**

# 《信息系统集成与开发》课程设计报告

## 问题背景

住房一直以来都是人民群众最为关心的问题，是关系民生的大事。但是过快上涨的商品房房价始终困扰着对住房与需求的人群，尤其是收入较低的人群，这已经发展成为了人民日常生活中的敏感问题，因此研究房地产价格的影响因素并对其进行预测有助于帮助对住房有需求的人民群众选择适当的买房时机，也有助于帮助政府相关人员了解房地产价格的走势并对其进行调控。

## 研究意义

衣食住行是人们生活的必需品，当房地产价格过高时，会影响人们的生活。此外，房地产行业在全球经济发展中发挥着重要作用。因此，深入研究影响商业地产价格的因素，预测商业地产价格的趋势，不仅有助于提高国家的竞争力，也有助于提高国家的竞争力。这不仅有助于国家调控商品房价格，使其在合理范围内波动，还能让人们在更合适的时间购买房屋。商业地产价格商业地产的价格受到许多因素的影响，如行政、经济、社会、位置等。这些因素的任何变化都会导致商业地产的价格变化。这些因素与商业房地产价格之间的关系是复杂和不确定的，因此有必要建立适当的模型来研究这种关系并预测商业房地产的价格。这些因素与商业地产价格之间的关系是复杂和不确定的，需要开发适当的模型来研究这种关系并预测价格趋势。

在对政府的实际影响方面，可以根据本文的结论制定一些房价政策。另一方面对开发商来说，它具有重要的经济意义。这是因为一个房地产项目从开始到结束涵盖了一个很长的时期，房价预测模型可以用来反映这一点。该项目实施时间较长，使用商品价格预测模型来捕捉一段时间内的价格变化，然后根据这些变化选择适当的投资方式和营销工具，这种模式对房地产开发商来说也非常有价值，因为它可以在一定程度上避免风险，获得相对更大的经济利益；对于那些希望购买或出售房地产的人来说，预测是非常有用的。对于那些想买房或卖房的人来说，房价预测模型可以预测商品房的价格，以便明智地计划买房或卖房，在一定程度上，它还可以消除风险，减少损失。

当房地产价格上涨时，人们不应惊慌失措，当房地产价格下跌时，也不应急于买房，而应根据实际情况正确、合理地买卖房屋。

## 研究目标与内容

本文的研究以贝叶斯算法为基础，对二手房价格进行分析和预测，对相关算法进行分析、比较、验证和检验，最后基于改进的贝叶斯算法对二手房价格进行分析和预测。相关的算法也将被分析、比较、验证和测试，最后将实现基于贝叶斯算法的房价分析和预测。

本文的主要工作是研究和调查一种基于数据分析的房屋价格预测方法，分析贝叶斯算法的原理和实现及其改进算法的原理与方法，并研究一个原型过滤系统的设计和实施。主要工作内容如下：

对现有过滤技术的发展状况，基本原理，方法和过程，进行分析和研究，主要过滤技术的优点和缺点比较。

根据分类后的二手房价格单位样本，对二手房价格信息进行分类检查，对每一类的二手房价格信息进行提取关键词，是二手房价格中的敏感符号，是表达该类别文本内容的类别名称，也就是分类向量。计算每种情况的出现的概率，就可以的到特征向量。

木文采用互信息的方法进行特征选取，计算每个关键词的RMI值，利用RMI值确定特征向量的分量词条。

解决基于向量空间模型的二手房价信息表示、数字分类中的特征向量的选择等问题，在核心的过滤实现技术中，重点是设计和实现一个基于贝叶斯算法的高级过滤系统，该算法基于一个简单的贝叶斯方法，完成二手房价信息分析及预测程。

## 设计报告结构

## 1 数据分析网站Weka

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis）是用于数据挖掘任务的机器学习算法的集合。它包含用于数据准备、分类、回归、聚类、关联规则挖掘和可视化的工具。同时Weka也是新西兰的一种鸟名，而Weka的主要开发者来自新西兰。

除了利用Weka进行数据挖掘的同时，Weka也很适合用来开发新的机器学习算法。[Weka是根据GNU通用公共许可证](https://www.gnu.org/licenses/gpl.html" \t "https://www.cs.waikato.ac.nz/ml/weka/_blank)发布的开源软件。Weka还支持[深度学习](https://deeplearning.cms.waikato.ac.nz/)。

如果想自己实现数据挖掘算法的话，可以参考Weka的接口文档。在Weka中集成自己的算法甚至借鉴它的方法自己实现可视化工具并不是件很困难的事情。

2005年8月，在第11届ACM SIGKDD国际会议上，怀卡托大学的Weka小组荣获了数据挖掘和知识探索领域的最高服务奖，Weka系统得到了广泛的认可，被誉为数据挖掘和机器学习历史上的里程碑，是现今最完备的数据挖掘工具之一（已有11年的发展历史）。Weka的每月下载次数已超过万次。

2014年3月起，[新西兰怀卡托大学](https://baike.baidu.com/item/%E6%96%B0%E8%A5%BF%E5%85%B0%E6%80%80%E5%8D%A1%E6%89%98%E5%A4%A7%E5%AD%A6/417707" \t "https://baike.baidu.com/item/weka/_blank)将推出Weka免费网课，课程分为初级和高级两个部分，每个部分时长5周。初级课程将于2014年3月3日开课，高级课程于2014年4月下旬开课。课程具体内容参见怀卡托大学网站Weka MOOC。课程在优酷网站也有专辑[1]。

## 2 Weka软件功能

Weka的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），同时Weka也是新西兰的一种鸟名，而Weka的主要开发者来自新西兰。Weka作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、[关联规则](https://baike.so.com/doc/6387919-6601574.html" \t "https://baike.so.com/doc/_blank)以及在新的交互式界面上的可视化[1]。

Weka系统汇集了最前沿的机器学习算法和数据预处理工具，以便用户能够快速灵活地将已有的处理方法应用于新的数据集。

2.1 Weka的历史

1992年末，新西兰怀卡托大学计算机科学系Ian Written博士申请基金。

1993年获新西兰政府资助，并于同年开发出接口和基础架构。

1994年发布了第一个Weka的内部版本。

1996年，第一个公开版本Weka2.1发布。Weka的早期版本主要采用C语言编写。

1997年，团队决定使用Java重新改写。Java允许用一个统一的接口来学习方案和方法的预处理和后处理，决定使用Java来代替C++或其他面向对象的语言，是因为Java编写的程序可以运行在绝大部分计算机上，而无须重新编译，更不需要修改源代码。Weka已经测试过的平台包括Linux,Windows和Macintosh操作系统，甚至PDA。最后的可执行程序复制过来即可运行，完全绿色，不需求复杂安装。当然，Java也有其缺点，最大的问题是它在速度上有缺陷，执行一个Java程序比相应的C语言程序马上好几倍。

1999年，发布Java的Weka3版本。

2013年，Weka的最新版本是3.7.8[2]。

2.2 处理方法

包括处理标准数据挖掘问题的所有方法：回归、分类、聚类、关联规则和属性选择。

2.2.1 分类与回归

Weka把分类（Classification）和回归（Regression）都放在“Classify”选项卡中，这是有原因的。

在这两个任务中，都有一个目标属性（输出变量）。我们希望根据一个样本（Weka中称作实例）的一组特征（输入变量），对目标进行预测。为了实现这一目的，我们需要有一个训练数据集，这个数据集中每个实例的输入和输出都是已知的。观察训练集中的实例，可以建立起预测的模型。有了这个模型，我们就可以新的输出未知的实例进行预测了。衡量模型的好坏就在于预测的准确程度。

在Weka中，待预测的目标（输出）被称作Class属性，这应该是来自分类任务的“类”。一般的，若Class属性是分类型时我们的任务才叫分类，Class属性是数值型时我们的任务叫回归。

2.2.2 聚类

聚类分析中的“类”（cluster）和前面分类的“类”（class）是不同的，对cluster更加准确的翻译应该是“簇”。聚类的任务是把所有的实例分配到若干的簇，使得同一个簇的实例聚集在一个簇中心的周围，它们之间距离的比较近；而不同簇实例之间的距离比较远。对于由数值型属性刻画的实例来说，这个距离通常指欧氏距离。

2.2.3 关联规则

关联规则又称购物栏分析。

目前，Weka的[关联规则](https://baike.so.com/doc/6387919-6601574.html" \t "https://baike.so.com/doc/_blank)分析功能仅能用来作示范，不适合用来挖掘大型数据集。

我们打算对前面的“bank-data”数据作关联规则的分析。用“Explorer”打开“bank-data-final.arff”后，切换到“Associate”选项卡。默认关联规则分析是用Apriori算法，我们就用这个算法，但是点“Choose”右边的文本框修改默认的参数，弹出的窗口中点“More”可以看到各参数的说明。

背景知识

首先我们来温习一下Apriori的有关知识。对于一条关联规则L->R，我们常用支持度（Support）和置信度（Confidence）来衡量它的重要性。规则的支持度是用来估计在一个购物栏中同时观察到L和R的概率P(L,R)，而规则的置信度是估计购物栏中出现了L时也出会现R的条件概率P(R|L)。[关联规则](https://baike.so.com/doc/6387919-6601574.html" \t "https://baike.so.com/doc/_blank)的目标一般是产生支持度和置信度都较高的规则[1]。

2.3 输入数据

通过以ARFF格式为代表的文件进行输入，然后直接读取数据库表。

2.3 Weka主界面Weka GUI Chooser

Weka主界面为Weka GUI选择器，它通过右边的四个按钮提供四种主要的应用程序供用户选择。

（1）探索者（Explorer），系统提供的最容易使用的图像用户接口。通过选择菜单和填写表单，可以调用Weka的所有功能。

虽然探索者界面使用很方便，但它也存在一个缺陷，要求它将所需数据全部一次读进内存，一旦用户打开某个数据集，就会读取全部数据。因此，这种批量方式仅适合处理中小规模的问题。而知识流刚好能够弥补这一缺陷。

（2）知识流（KnowledgeFlow），使用增量方式的算法来处理大型数据集，用户可以定制处理数据流的方式和顺序。按照一定顺序将代表数据源、预处理工具、学习算法、评估手段和可视化模块的各构件组合在一起，形成数据流。

（3）实验者（Experimenter），用于帮助用户解答实际应用分类和回归技术中遇到的一个基本问题——对于一个已知问题，哪种方法及参数值能够取得最佳效果？尽管探索者界面也能通过交互完成这样的功能，但通过实验者界面，用户可以让处理过程实现自动化。实验者界面更加容易使用不同参数去设置分类器和过滤器，使之运行在一组数据集中，收集性能统计数据，实现重要的测试实验。

（4）简单命令行（Simple CLI），这个界面是为不提供自己的命令行界面的操作系统提供的，该简单命令行界面用户和用户进行交互，可以直接执行Weka命令[2]。

## 3 数据集合定义

3.1 数据集合格式

跟很多[电子表格](https://baike.baidu.com/item/%E7%94%B5%E5%AD%90%E8%A1%A8%E6%A0%BC" \t "https://baike.baidu.com/item/weka/_blank)或数据分析软件一样，Weka所处理的数据集是一个二维的表格。

这里我们要介绍一下Weka中的术语。表格里的一个横行称作一个实例（Instance），相当于统计学中的一个样本，或者数据库中的一条记录。竖行称作一个属性（Attribute），相当于统计学中的一个变量，或者数据库中的一个字段。这样一个表格，或者叫数据集，在Weka看来，呈现了属性之间的一种关系（Relation）。示例表格一共有14个实例，5个属性，关系名称为“weather”。

Weka存储数据的格式是ARFF（Attribute-Relation File Format）文件，这是一种ASCII文本文件。[二维表格](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4%E8%A1%A8%E6%A0%BC" \t "https://baike.baidu.com/item/weka/_blank)存储在如下的ARFF文件中。这也就是Weka自带的“weather.arff”文件，在Weka[安装目录](https://baike.baidu.com/item/%E5%AE%89%E8%A3%85%E7%9B%AE%E5%BD%95" \t "https://baike.baidu.com/item/weka/_blank)的“data”子目录下可以找到[1]。

下面我们来对这个文件的内容进行说明。

识别ARFF文件的重要依据是分行，因此不能在这种文件里随意的断行。空行（或全是空格的行）将被忽略。

以“%”开始的行是注释，Weka将忽略这些行。如果你看到的“weather.nominal.arff”文件多了或少了些“%”开始的行，是没有影响的[4]。

除去注释后，整个ARFF文件可以分为两个部分。第一部分给出了头信息（Head information），包括了对关系的声明和对属性的声明。第二部分给出了数据信息（Data information），即数据集中给出的数据。从“@data”标记开始，后面的就是数据信息了[4]。

3.1.1 关系声明

关系名称在arff文件的第一个有效行来定义，格式为：

@relation <relation-name>

<relation-name>是一个字符串。如果这个字符串包含空格，它必须加上引号（指英文标点的单引号或双引号）[4]。

3.1.2 属性声明

属性声明用一列以“@attribute”开头的语句表示。数据集中的每一个属性都有它对应的“@attribute”语句，来定义它的属性名称和数据类型。

这些声明语句的顺序很重要。首先它表明了该项属性在数据部分的位置。例如，“humidity”是第三个被声明的属性，这说明数据部分那些被逗号分开的列中，第三列数据 85 90 86 96 ... 是相应的“humidity”值。其次，最后一个声明的属性被称作class属性，在分类或回归任务中，它是默认的目标变量。

属性声明的格式为：@attribute <attribute-name> <datatype>

其中<attribute-name>是必须以字母开头的字符串。和关系名称一样，如果这个字符串包含空格，它必须加上引号。

Weka支持的<datatype>有四种，分别是：numeric-------------------------数值型、<nominal-specification>-----分类（nominal）型、string----------------------------字符串型、date [<date-format>]--------日期和时间型[4]。

其中<nominal-specification> 和<date-format> 将在下面说明。还可以使用两个类型“integer”和“real”，但是Weka把它们都当作“numeric”看待。注意“integer”，“real”，“numeric”，“date”，“string”这些关键字是区分大小写的，而“relation”“attribute”和“date”则不区分[4]。

（1）数值属性

数值型属性可以是整数或者实数，但Weka把它们都当作实数看待。

（2）分类属性

分类属性由<nominal-specification>列出一系列可能的类别名称并放在花括号中：{<nominal-name1>, <nominal-name2>, <nominal-name3>, ...} 。数据集中该属性的值只能是其中一种类别。

例如如下的属性声明说明“outlook”属性有三种类别：“sunny”，“ overcast”和“rainy”。而数据集中每个实例对应的“outlook”值必是这三者之一。

@attribute outlook {sunny, overcast, rainy}，如果类别名称带有空格，仍需要将之放入引号中。

（3）字符串属性

字符串属性中可以包含任意的文本。这种类型的属性在文本挖掘中非常有用。

示例：@ATTRIBUTE LCC string

（4）日期和时间属性

日期和时间属性统一用“date”类型表示，它的格式是：@attribute <name> date [<date-format>]

其中<name>是这个属性的名称，<date-format>是一个字符串，来规定该怎样解析和显示日期或时间的格式，默认的字符串是ISO-8601所给的日期时间组合格式“yyyy-MM-ddTHH:mm:ss”[4]。

数据信息部分表达日期的字符串必须符合声明中规定的格式要求（下文有例子）。

3.1.3 数据信息

数据信息中“@data”标记独占一行，剩下的是各个实例的数据。

每个实例占一行。实例的各属性值用逗号“,”隔开。如果某个属性的值是缺失值（missing value），用问号“?”表示，且这个问号不能省略。例如：

@data

sunny,85,85,FALSE,no

?,78,90,?,yes

字符串属性和分类属性的值是区分大小写的。若值中含有空格，必须被引号括起来。[4]例如：

@relation LCCvsLCSH

@attribute LCC string

@attribute LCSH string

@data

AG5, 'Encyclopedias and dictionaries.;Twentieth century.'

AS262, 'Science -- Soviet Union -- History.'

日期属性的值必须与属性声明中给定的相一致。例如：

@RELATION Timestamps

@ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"

@DATA

"2001-04-03 12:12:12"

"2001-05-03 12:59:55"

3.1.4 稀疏数据

有的时候数据集中含有大量的0值（比如购物篮分析），这个时候用稀疏格式的数据存贮更加省空间。

稀疏格式是针对数据信息中某个实例的表示而言，不需要修改ARFF文件的其它部分。看如下的数据：

@data

0, X, 0, Y, "class A"

0, 0, W, 0, "class B"

用稀疏格式表达的话就是：

@data

{1 X, 3 Y, 4 "class A"}

{2 W, 4 "class B"}

每个实例用花括号括起来。实例中每一个非0的属性值用<index> <空格> <value>表示。<index>是属性的序号，从0开始计；<value>是属性值。属性值之间仍用逗号隔开。这里每个实例的数值必须按属性的顺序来写，如：{1 X, 3 Y, 4 "class A"}，不能写成{3 Y, 1 X, 4 "class A"}。

注意在稀疏格式中没有注明的属性值不是缺失值，而是0值。若要表示缺失值必须显式的用问号表示出来[4]。

3.1.5 Relational型属性

在Weka 3.5版中增加了一种属性类型叫做Relational，有了这种类型我们可以像关系型数据库那样处理多个维度了。但是这种类型目前还不见广泛应用，暂不作介绍[4]。

3.2 csv转换为arff格式

我们通常接触到的数据为csv格式较多，若想要将数据更改为Weka通用的arff格式，我们可以利用Weka工具进行格式的转换。

3.2.1 csv格式

逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）。纯文本意味着该文件是一个字符序列，不含必须像二进制数字那样被解读的数据。CSV文件由任意数目的记录组成，记录间以某种换行符分隔；每条记录由字段组成，字段间的分隔符是其它字符或字符串，最常见的是逗号或制表符。通常，所有记录都有完全相同的字段序列。通常都是纯文本文件。建议使用Word或是记事本来开启，再则先另存新档后用Excel开启，也是方法之一[6]。

3.2.2 arff格式

arff是一种Weka专用的文件格式，由Andrew Donkin创立，有传言说arff代表Andrew's Ridiculous File Format(安德鲁的荒唐文件格式)，但在Weka的正式文档中明确说明arff代表Attribute-Relation File Format(属性——关系文件格式)。该文件是ASCII文本文件，描述共享一组属性结构的实例列表，由独立且无序的实例组成，是Weka表示数据集的标准方法，arff不涉及实例之间的关系。

3.2.3 获取csv文件

本次案例我们使用UCI网站中的Soybean(Large)数据进行演示，我们先从网站中下载下来我们所需要的数据，进入网站后选择Download按钮进入下载页面，我们选择adult.data数据进行下载下来发现是一个纯文本文件，用记事本打开，发现此文件大约有32561多条实例，属性多达15个，为了方便用表展示，我们只截取了前5个，实例也仅选取了前10个以作示范。

将网站主页中的属性名依次复制到该data文件中，属性之间用逗号隔开，这样我们就构造出来了csv格式的数据文件。保存文件，将该文件的后缀改为csv，我们就得到这个数据集的csv格式文件。

表3-1 Adult数据集合

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 索引值 | age | workclass | fnlwgt: | education | education-num |
| 1 | 39 | State-gov | 77516 | Bachelors | 13 |
| 2 | 50 | Self-emp-not-inc | 83311 | Bachelors | 13 |
| 3 | 38 | Private | 215646 | HS-grad | 9 |
| 4 | 53 | Private | 234721 | 11th | 7 |
| 5 | 28 | Private | 338409 | Bachelors | 13 |
| 6 | 37 | Private | 284582 | Masters | 14 |
| 7 | 49 | Private | 160187 | 9th | 5 |
| 8 | 52 | Self-emp-not-inc | 209642 | HS-grad | 9 |
| 9 | 31 | Private | 45781 | Masters | 14 |
| 10 | 42 | Private | 159449 | Bachelors | 13 |

3.2.4 转换csv格式为arff格式

1.打开Weka，选择explorer模式

2.这时会打开新的explorer窗口，我们选择Open File选项，可以对我们的数据进行选择。

3.打开文件后，我们点击save选择文件类型为arff，选择文件名称以及存放位置后，点击保存，这样我们的数据就保存为了arff格式。

3.3 UCI数据集的介绍

UCI（University of California Irvine）数据集是美国加州大学欧文分校提出的一种适合模式识别和机器学习方向的开源数据集，很多学者选择使用UCI上的数据集来验证自己所提算法的正确性。博文写作时已拥有488个数据集，数据集还在不断扩充中，这些数据集主要分为二值分类问题、多分类问题以及回归拟合问题。UCI数据集提供了各个数据集的上主要属性，可以根据自己提出的各类算法在其数据集上做实验结果论证，证明自己所提算法的合理性[3]。

3.4 数据集合转换

使用Weka作数据挖掘，面临的第一个问题往往是我们的数据不是ARFF格式的。幸好，Weka还提供了对CSV文件的支持，而这种格式是被很多其他软件所支持的。此外，Weka还提供了通过JDBC访问数据库的功能[5]。

首先，我们从UCI数据集网站（https://archive-beta.ics.uci.edu/）上获取csv文件，以Iris数据集为例，在网站上下载iris.date文件和iris.names文件，然后将下载好的iris.date文件后缀名改为.csv格式。然后点开查看发现全是各种数据，但是没有相应的描述，这时候以记事本打开iris.names文件找到属性信息（Attribute Information），手工添加一行属性名，添加好后保存。注意属性名的个数要跟数据属性的个数一致，仍用逗号隔开。注意，Weka必须从CSV文件的第一行读取属性名，否则就会把第一行的各属性值读成变量名。

打开Weka，点开Explorer页面选择open file，选择显示.csv文件,选择刚刚改好的iris.csv文件就可以打开了。如果想转换为ARFF类型，就点击保存（save），保存为.arff文件就可以了。

## 参考文献

1. 百度百科, URL: https://baike.baidu.com/item/weka/10701215?fr=aladdin. [2022-6-16].

[2] zuokankan, URl: http://t.zoukankan.com/chamie-p-4551514.html. [2022-6-16].

[3] CSDN, URL: <https://blog.csdn.net/qq_32892383/article/details/104424358.> [2022-6-17]

[4] 博客园, URL: https://www.cnblogs.com/xiami303/articles/2288119.html. [2022-6-18].

[5] 博客园, URL: https://www.cnblogs.com/tanrong/p/8671068.html. [2022-6-18].

[6] 博客, URL: https://baike.baidu.com/item/CSV/10739. [2022-6-18].