北京交通大学海滨学院

课 程 设 计 报 告

《信息系统集成与开发》

姓　　名： 娄建鹏

学　　号： 19851002

专　　业： 计算机科学与技术

系　　别： 计算机与信息技术

指导教师： 王志海

2022 年 6 月

中国联通套餐

此次分析报告我将使用中国联通套餐的数据集进行分析，通过大量数据分析用户想要那种套餐，分析软件使用 Weka 进行分析。

首先我们先要对数据进行预处理，剔除一些空数据、重复以及无效数据；后我将用随机森林模型和贝叶斯模型分析数据，之后我们把两者进行对比分析，最后，预测分析出用户消费习惯及偏好，匹配用户最合适的套餐，提升用户感知，带动用户需求，从而帮助用户发现合适套餐，也能将合适套餐信息推送给用户。

1 数据分析网站简介

本小节主要介绍了 Weka 网站的基本内容，项目目标及其本报告的组织结构。

1.1 Weka 网站介绍

Weka 是新西兰怀卡托大学用 Java 开发的数据挖掘著名开源软件，该系统自 1993 年开始由新西兰政府资助，至今已经历了 20 年的发展，其功能已经十分强大和成熟。Weka 集合了大量的机器学习和相关技术，受领域发展和用户需求所推动，代表了当今数据挖掘和机器学习领域的最高水平。

Weka 全名是怀卡托智能分析环境，非商业化的，基于 Java 环境下开源的机器学习，是收集数据采集任务的机器学习算法。它含有用于数据准备、分类、回归、聚类、关联规则采集和可视化的工具。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。

Weka 仅被发现于新西兰的岛屿上，Weka 是不会飞的鸟类，他具有好奇的特性。这个名字的发音像这样，并且鸟儿的叫声听起来也像这样。

Weka 是在 GNU 通用公共许可证的发布下的开源软件。并且，我们放了一些免费课程在网上，教授使用 Weka 用来机器学习和数据采集，这些视频课程可以在 Youtube 上找到。而且，Weka 支持深度学习。

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。

Weka 的主要开发者同时恰好来自新西兰的怀卡托大学。

1.2 Weka 项目目标

第一，使 ML 技术普遍可用；第二，将其应用于对新西兰工业至关重要的实际问题；第三，开发新的机器学习算法并向世界推广；最后，为该领域的理论框架做出贡献。

1.3 课程设计报告组织结构

本报告主要针对于中国联通套餐数据集的，利用 Weka 软件进行讲究与预测，从而实现中国联通套餐的选择。

全文共 6 节，每一节内容如下：

第 1 节　数据分析网站简介：介绍了 Weka 网站的基本内容，项目目标及其本报告的组织结构。

第 2 节　Weka 软件系统功能及其概念：介绍了机器学习与数据挖掘的最基本的概念，以及 Weka 的基本结构与功能和 Weka 提供的四个功能环境。

第 3 节　Weka 数据集合格式及其转换方法：介绍了 Weka 数据集合的格式，把其他格式的文件转换成 arff 文件，并且举了一个例子进行说明。

第 4 节　中国联通套餐的数据集来源与处理：介绍了中国联通套餐的数据集来源，中国联通套餐背景及其目的，数据展示，数据属性，数据预处理。

第 5 节　中国联通套餐的分析和预测及可视化：介绍了如何将中国联通套餐的数据集转换为 arff 格式，算法模型评估，随机森林模型介绍，贝叶斯模型介绍，两种模型的对比，以及根据这个联通套餐数据集最后得出的结论。

2　Weka 软件系统功能及其概念

本小节主要介绍了机器学习与数据挖掘的最基本的概念，以及 Weka 的基本结构与功能和 Weka 提供的四个功能环境。

2.1　Weka 功能介绍

Weka 是一种使用 Java 语言编写的数据挖掘机器学习软件，是 GNU 协议下分发的开源软件。主要用于科研、教育和应用领域。并且，Weka 系统汇集了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成熟处理方法应用于新的数据集。其功能如下。

2.1.1　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。

2.1.2　输入数据

一种是直接通过以 ARFF 格式为代表的文件进行输入，二是直接读取数据表。

2.1.3　界面功能

在 Weka 内进行的是数据预处理，训练，验证这三个步骤。数据预处理包括特征选择，特征值处理（比如归一化），样本选择等操作。训练包括算法选择，参数调整，模型训练。验证是对模型结果进行验证。数据预处理，打开 Explorer 界面，点”open file”，在 Weka 安装目录下，选择 data 目录里的”labor.arff”文件，会看到主

界面，将整个区域分为七个部分。在 Explorer 中，打开 classifer 选项卡，整个界面被分成几个区域。分别是 Classifier，点击 choose 按钮，可以选择 Weka 提供的分类器。Test options 评价模型效果的方法，有四个选项。可视化，打开 Explorer 的 Visualize 面板，可以看到最上面是一个二维的图形矩阵，该矩阵的行和列均为所有的特征（包括类别标签），第 i 行第 j 列表示特征 i 和特征 j 在二维平面上的分布情况。图形上的每个点表示一个样本，不同的类别使用不同的颜色标识。

2.2 Weka 软件系统的四个功能环境

Weka 的窗口右侧共有四个应用，Explorer（探索者）界面，是 Weka 的主要图形化用户界面，其全部功能都可通过菜单选择或表单填写进行访问,用来进行数据实验、挖掘的环境，它提供了分类，聚类，关联规则，特征选择，数据可视化的功能。Experimentor 用来进行实验，对不同学习方案进行数据测试的环境。KnowledgeFlow 功能和 Explorer 差不多，不过提供的接口不同，用户可以使用拖拽的方式去建立实验方案。另外，它支持增量学习。SimpleCLI 简单的命令行界面。

2.2.1 简单 CLI（SimpleCLI）

提供了一个简单的命令行界面，从而可以在没有自带命令行的操作系统中直接 i 执行 Weka 命令。

使用命令行有两个好处：一个是可以把模型保存下来，这样有新的待预测数据出现时，不用每次重新建模，直接应用保存好的模型即可。另一个是对预测结果给出了置信度，我们可以有选择的采纳预测结果，例如，只考虑那些置信度在 85%以上的结果。

2.2.2 探索者（Explorer）

使用 Weka 探索数据的环境。在这个环境中，Weka 提供了数据的预处理，数据格式的转化（从 CSV 格式到 ARFF 格式的转化，详见第 4 部分），各种数据挖掘算法（包括分类与回归算法，聚类算法，关联规则等），并提供了结果的可视化工具。

对于一个数据集，通过简单的数据的预处理，并对数据挖掘算法进行选择（在Weka3.5 版本之后，加入了算法的过滤功能，可以过滤掉那些不适合当前数据集类型的算法），接着通过窗口界面对算法的参数进行配置，最后点击“Start”按钮就可以运行了。

可视化工具分为对数据集的可视化和对部分结果的可视化，并且我们可以通过属性选择工具(Select Attribute)，通过搜索数据集中所有属性的可能组合，找出预测效果最好的那一组属性。

2.2.3 实验者（Experimenter）

运行算法试验、管理算法方案之间的统计检验的环境。Experiment 环境可以让用户创建，运行，修改和分析算法试验，这也许比单独的分析各个算法更加方便。例如，用户可创建一次试验，在一系列数据集上运行多个算法（schemes），然后分析结果以判断是否某个算法比其他算法（在统计意义下）更好。

Explorermenter 主要包括简单模式，复杂模式和远程模式。复杂模式是对简单模式的基本功能的扩充，而远程模式允许我们通过分布式的方法进行实验。
就功能模块而言，分为设置模块，运行模块和分析模块。在设置模块中我们可以自定义实验，加入多个算法和多方的源数据（支持 ARFF 文件，CSV 文件和数据库），在运行模块中我们可以运行我们的实验，而在分析模块中，我们可以分析各种算法的准确性，并提供了各种统计方法对结果进行检验比较。

值得一提的是，我们可以把实验的各种参数，包括算法，数据集等，保存以方便下一次相同实验的进行；也可以把各种算法保存，方便应用在不同的数据集上；如果数据集来源于数据库的话，实验在过程中可以中止并继续（原因可以是被中止或者是扩展了实验），这样就不用重新运行那些已实验过的算法/数据集祝贺，而仅计算还没有被实验的那些。

2.2.4 知识流（KnowledgeFlow）

这个环境本质上和 Explorer 所支持的功能是一样的，但是它有一个可以拖放的界面。它有一个优势，就是支持增量学习（incremental learning）。KnowledgeFlow 为 Weka 提供了一个 " 数据流 " 形式的界面．用户可以从一个工具栏中选择组件，

把它们放置在面板上并按一定的顺序连接起来，这样组成一个＂知识流＂（knowledge flow）来处理和分析数据．目前，所有的 Weka 分类器（classifier）、筛选器（filter）、聚类器（clusterer）、载入器（loader）、保存器（saver），以及一些其他的功能可以在 KnowledgeFlow 中使用。

KnowledgeFlow 可以使用增量模式（incrementally）或者批量模式（inbatches）来处理数据（Explorer 只能使用批量模式）。当然对数据进行增量学习要求分类器能够根据各实例逐个逐个的更新．现在 Weka 中有五个分类器能够增量地处理数据：NaiveBayesUpdateable，IB1，IBk，LWR（局部加权回归）。还有一个 meta 分类器 RacedIncrementalLogitBoost 可以使用任意基于回归的学习器来增量地学习离散的分类任务。

2.3 Weka 的基本结构的概念

Weka 是收集数据采集任务的机器学习算法。它含有用于数据准备、分类、回归、聚类、关联规则采集和可视化的工具，其基本结构如下。

2.3.1 关联规则

关联规则又称购物栏分析。目前，Weka 的关联规则分析功能仅能用来作示范，不适合用来挖掘大型数据集。

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。

2.3.2 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。

2.3.3 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。聚类分析，通常使用最常见的 K 均值（K-means）算法。

2.3.4 数据准备

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。

2.3.5 可视化

Weka 系统汇集了前沿的数据预处理工具，以便用户能够快速灵活地将已有的数据数据处理成可视化图，方便用户可以直接的帮助人更好的分析数据。

2.4 机器学习的分类

机器学习分为两种主要类型其分类为。

2.4.1 有监督学习（预测学习，分类学习）

目标是在给定一系列输入/输出实例所构成的数据集的条件下，学习输入 x 到输出 y 的映射关系。这里的数据集称为训练集，实例的个数称为训练样本数。（从分类角度看，即训练集中各组数据的类别已知）可以对所观察到的值与预测值进行比较，得到明确的误差值。

2.4.2 无监督学习（描述学习）

目标是在给定一系列输入实例构成的数据集的条件下，发现数据中的有趣模式。因为我们不知道需要寻找什么样的模式，也没有明显的误差度量可供使用。

2.5 数据挖掘

数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

数据挖掘是人工智能和数据库领域研究的热点问题，所谓数据挖掘是指从数据库的大量数据中揭示出隐含的、先前未知的并有潜在价值的信息的非平凡过程。数据挖掘是一种决策支持过程，它主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据，作出归纳性的推理，从中挖掘出潜在的模式，帮助决策者调整市场策略，减少风险，作出正确的决策。知识发现过程由以下三个阶段组成：①数据准备；②数据挖掘；③结果表达和解释。数据挖掘可以与用户或知识库交互。

数据挖掘是通过分析每个数据，从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示三个步骤。数据准备是从相关的数据源中选取所需的数据并整合成用于数据挖掘的数据集；规律寻找是用某种方法将数据集所含的规律找出来；规律表示是尽可能以用户可理解的方式（如可视化）将找出的规律表示出来。数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异群组分析和演变分析等。

2.5.1 数据挖掘的步骤

数据挖掘过程模型步骤主要包括定义问题、建立数据挖掘库、分析数据、准备数据、建立模型、评价模型和实施。下面让我们来具体看一下每个步骤的具体内容：

(1)定义问题。在开始知识发现之前最先的也是最重要的要求就是了解数据和业务问题。必须要对目标有一个清晰明确的定义，即决定到底想干什么。比如，

想提高电子信箱的利用率时，想做的可能是“提高用户使用率”，也可能是“提高一次用户使用的价值”，要解决这两个问题而建立的模型几乎是完全不同的，必须做出决定。

(2)建立数据挖掘库。建立数据挖掘库包括以下几个步骤：数据收集，数据描述，选择，数据质量评估和数据清理，合并与整合，构建元数据，加载数据挖掘库，维护数据挖掘库。

(3)分析数据。分析的目的是找到对预测输出影响最大的数据字段，和决定是否需要定义导出字段。如果数据集包含成百上千的字段，那么浏览分析这些数据将是一件非常耗时和累人的事情，这时需要选择一个具有好的界面和功能强大的工具软件来协助你完成这些事情。

(4)准备数据。这是建立模型之前的最后一步数据准备工作。可以把此步骤分为四个部分：选择变量，选择记录，创建新变量，转换变量。

(5)建立模型。建立模型是一个反复的过程。需要仔细考察不同的模型以判断哪个模型对面对的商业问题最有用。先用一部分数据建立模型，然后再用剩下的数据来测试和验证这个得到的模型。有时还有第三个数据集，称为验证集，因为测试集可能受模型的特性的影响，这时需要一个独立的数据集来验证模型的准确性。训练和测试数据挖掘模型需要把数据至少分成两个部分，一个用于模型训练，另一个用于模型测试。

(6)评价模型。模型建立好之后，必须评价得到的结果、解释模型的价值。从测试集中得到的准确率只对用于建立模型的数据有意义。在实际应用中，需要进一步了解错误的类型和由此带来的相关费用的多少。经验证明，有效的模型并不一定是正确的模型。造成这一点的直接原因就是模型建立中隐含的各种假定，因此，直接在现实世界中测试模型很重要。先在小范围内应用，取得测试数据，觉得满意之后再向大范围推广。

(7)实施。模型建立并经验证之后，可以有两种主要的使用方法。第一种是提供给分析人员做参考；另一种是把此模型应用到不同的数据集上。

2.5.2 数据挖掘分析方法

数据挖掘分为有指导的数据挖掘和无指导的数据挖掘。有指导的数据挖掘是利用可用的数据建立一个模型，这个模型是对一个特定属性的描述。无指导的数据挖掘是在所有的属性中寻找某种关系。具体而言，分类、估值和预测属于有指导的数据挖掘；关联规则和聚类属于无指导的数据挖掘。

1　分类。它首先从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘技术，建立一个分类模型，再将该模型用于对没有分类的数据进行分类。

2　估值。估值与分类类似，但估值最终的输出结果是连续型的数值，估值的量并非预先确定。估值可以作为分类的准备工作。

3　预测。它是通过分类或估值来进行，通过分类或估值的训练得出一个模型，如果对于检验样本组而言该模型具有较高的准确率，可将该模型用于对新样本的未知变量进行预测。

4　相关性分组或关联规则。其目的是发现哪些事情总是一起发生。

5　聚类。它是自动寻找并建立分组规则的方法，它通过判断样本之间的相似性，把相似样本划分在一个簇中。

3　Weka 数据集合格式及其转换方法

本小节主要介绍了 Weka 数据集合的格式，把其他格式的文件转换成 arff 文件，并且举了一个例子来进行说明。

3.1　Weka 数据集合格式

跟很多电子表格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)。

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。也可以直接使用 csv 文件格式的文件，但与传统 csv 文件不同，Weka 能识别的 csv 文件要求第一行给各列的定义。因为 csv 文件比较容易获得，excel 表格文件可以直接另存为 csv 文件。推荐使用 csv 文件。二维表格存储在如下的 arff 文件中。这也就是 Weka 自带的“weather.arff”文件，在 Weka 安装目录的“data”子目录下可以找到见表 3-1。

表 3-1　Weather 数据集合

序号	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	No

3.2　将其他格式转换为 arff 格式

Weka 支持很多种文件格式，包括 arff、xrff、csv，甚至有 libsvm 的格式。其中，arff 是最常用的格式。而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[4]。

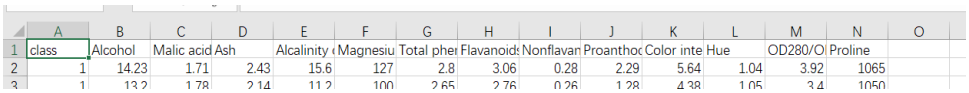
3.2.2 arff 格式

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 例子和过程

首先，本次案例我们使用 UCI 网站中的 Wine 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 后的 Data Folder 按钮进入下载页面，我们选择 wine.data 和 wine.names 数据和进行下载下来发现是 2 个纯文本文件。

然后，把 wine.data 文件的后缀改成 csv。双击打开，并在第一行上边添加一行，再将 wine.names 文件打开，把 The attributes are 下的名字依次填入 wine.csv 文件上面空出来的一行中。第一列是 class。然后保存，我们就得到这个数据集的 csv 格式文件见图 3-1。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	class	Alcohol	Malic acid	Ash	Alcalinity	Magnesiu	Total phe	Flavanoid	Nonflavan	Proanthoc	Color inte	Hue	OD280/O	Proline	
2	1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065	
3	1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050	

图 3-1　wine.data 命名

之后打开 Weka，选择 explorer 模式，这时会打开新的 explorer 窗口，我们点击左上角 Open File 选项，在 Look In 中找到你存放 wine.csv 的位置，并且把底下的 File of Type 改成 CSV data files 格式，找到文件并打开见图 3-2。

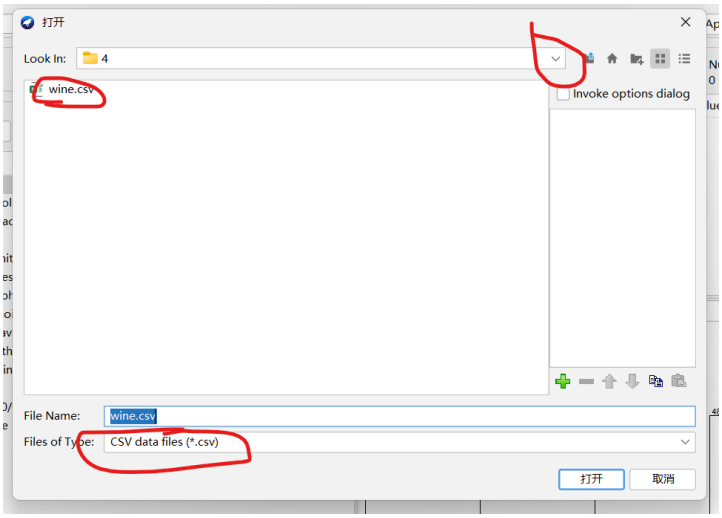


图 3-2　如何找到 wine.csv 文件

最后，点击右上角的 Save…按钮，把 File Name 中的 wine.csv 的后缀改成 arff，然后再点击保存，这样我们的数据就保存为了 arff 格式见图 3-3。

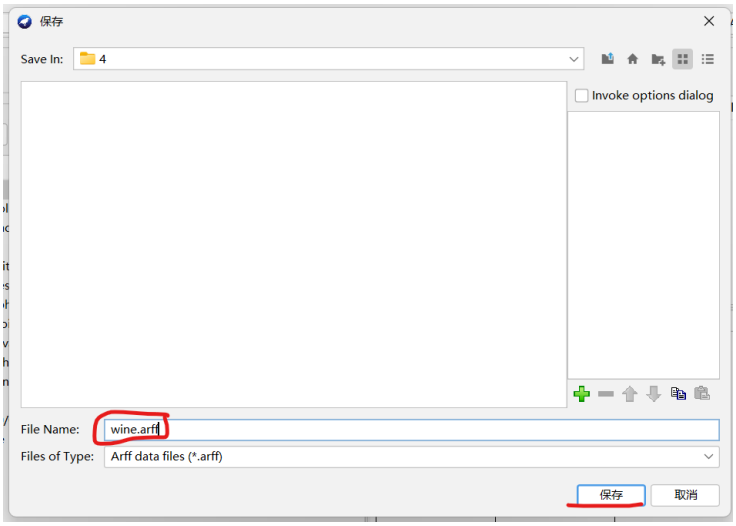


图 3-3　如何保存成 arff 文件

由于此文件大约有 170 多条实例，属性多达 13 个，而表不好展示，为了方便说明我们分条来说。

（1）class：种类，数值类型，表示红酒的类型，范围在 1-3。

（2）Alcohol：酒精含量，数值类型，表示红酒的度数，范围在 10-15°。

（3）Malic acid：苹果酸含量，数值类型，表示红酒的苹果酸含量，范围在 0-6。

（4）Ash：灰分，数值类型，表示红酒中残留物含量，范围在 1-4mg。

（5）Alcalinity of ash：灰分的碱性，数值类型，表示红酒中灰分的 ph 值，范围在 12-30。

（6）Magnesium：镁含量，数值类型，表示红酒中镁的含量，范围在 6-200mg。

（7）Total phenols：总酚类，数值类型，表示红酒中总酚的含量，范围在 0-5mg。

（8）Flavanoids：类黄烷，数值类型，表示红酒中类黄烷的含量，范围在 0-5mg。

（9）Nonflavanoid phenols：非类黄酮酚，数值类型，表示红酒中非类黄酮酚的含量，范围在 0-1mg。

（10）Proanthocyanins：原花青素，数值类型，表示红酒中原花青素的含量，范围在 0-3mg。

（11）Color intensity：颜色强度，数值类型，表示红酒的颜色深浅，范围在 1-15。

（12）Hue：色调，数值类型，表示红酒的色调，范围在 0-2。

（13）OD280/OD315 of diluted wines：稀释葡萄酒的 OD280/OD315，数值类型，范围在 1-5。

（14）Proline：脯氨酸，数值类型，表示红酒中脯氨酸的含量，范围在 300-1500mg。

部分数据展示见图 3-4

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	class	Alcohol	Malic acid	Ash	Alcalinity o	Magnesiu	Total phe	Flavanoid	Nonflavan	Proanthoc	Color inte	Hue	OD280/O	Proline	
2	1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065	
3	1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050	
4	1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185	
5	1	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480	
6	1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735	
7	1	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450	
8	1	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290	
9	1	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295	
10	1	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045	
11	1	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045	
12	1	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510	
13	1	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280	
14	1	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320	
15	1	14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150	
16	1	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547	
17	1	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310	
18	1	14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280	
19	1	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130	
20	1	14.19	1.59	2.48	16.5	108	3.3	3.93	0.32	1.86	8.7	1.23	2.82	1680	

图 3-4　wine 数据展示

4 中国联通套餐的数据集来源与处理

本小节主要介绍了中国联通套餐的数据集来源，中国联通套餐背景及其目的，数据展示，数据属性，数据预处理。

4.1 中国联通套餐的数据集来源

中国联通套餐的数据集选自 datafountain 网站，其中，共有两个数据集。一个是测试集：共包含 160567 个用户选择的套餐信息。另一个是训练集：共包含 374656 个用户选择的套餐信息。这两个数据集的信息都是真实用户选择的套餐信息。该网站的目标是构建中国最有影响力和权威度的数据科学与大数据分析处理竞赛平台，平台的宗旨是“数据互联、大众创新”，是大数据资源和需求的汇聚地，是优秀数据科学家的俱乐部，是“大众创业，万众创新”的在线空间。

由于我这个题目在 datafountain 网站需要报名才能下载，所以我根据这个竞赛的名字在网上找到了数据集的下载地址。

数据集的下载网址为 https://zhuanlan.zhihu.com/p/51782481。

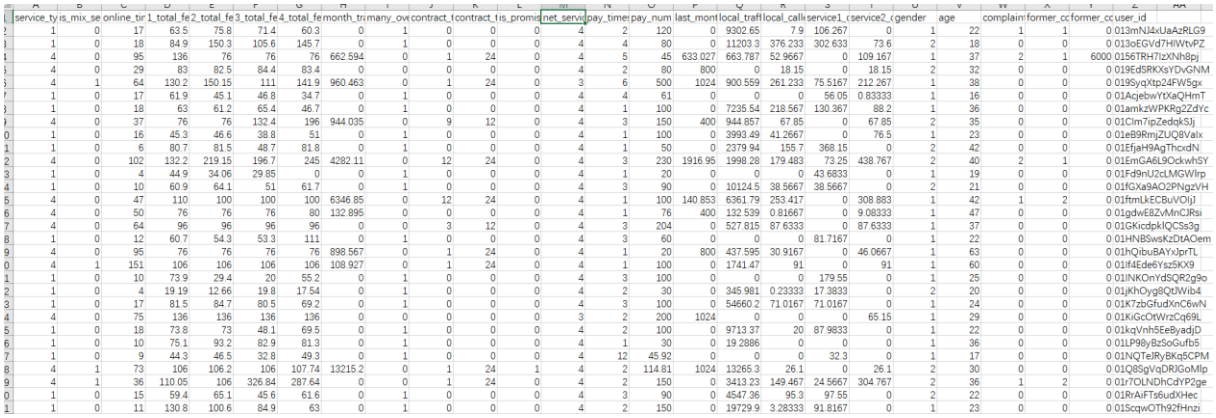
4.2 中国联通套餐背景及目的

联通产业作为国家基础产业之一，覆盖广、用户多，在支撑国家建设和发展方面尤为重要。随着互联网技术的快速发展和普及，用户消耗的流量也成井喷态势，近年来，联通运营商推出大量的联通套餐用以满足用户的差异化需求，面对种类繁多的套餐，如何选择最合适的一款对于运营商和用户来说都至关重要，尤其是在联通市场增速放缓，存量用户争夺愈发激烈的大背景下。针对联通套餐的个性化推荐问题，通过数据挖掘技术构建了基于用户消费行为的联通套餐个性化推荐模型，根据用户业务行为画像结果，分析出用户消费习惯及偏好，匹配用户最合适的套餐，提升用户感知，带动用户需求，从而达到用户价值提升的目标。

套餐的个性化推荐，能够在信息过载的环境中帮助用户发现合适套餐，也能将合适套餐信息推送给用户。各种套餐满足了用户有明确目的时的主动查找需求，而个性化推荐能够在用户没有明确目的的时候帮助他们发现感兴趣的新内容。

4.3 数据展示

由于中国联通套餐的数据集有两个，但是两个数据集里的属性基本都一样，所以我们只展示测试集的部分数据见图 4-1。



service_ty	is_mix_se	online_tir	1_total_fe	2_total_fe	3_total_fe	4_total_fe	month_tri	many_ov	contract_t	contract_t	is_promis	net_servic	pay_time	pay_num	last_mont	local_traff	local_call	service1_	service2_	gender	age	complain	former_cc	former_cc	user_id
1	0	17	63.5	75.8	71.4	60.3	0	1	0	0	0	4	2	120	0	9302.65	7.9	106.267	0	1	22	1	1	0	013mNJ4xUaAzRLG9
1	0	18	84.9	150.3	105.6	145.7	0	1	0	0	0	4	4	80	0	11203.3	376.233	302.633	73.6	2	18	0	0	0	013oEGVd7HlWtvPZ
4	0	95	136	76	76	76	662.594	0	1	24	0	4	5	45	633.027	663.787	52.9667	0	109.167	1	37	2	1	6000	0156TRH7IzXNh8pj
4	0	29	83	82.5	84.4	83.4	0	0	0	0	0	4	2	80	800	0	18.15	0	18.15	2	32	0	0	0	019EdSRKXsYDvGNM
4	1	64	130.2	150.15	111	141.9	960.463	0	1	24	0	3	6	500	1024	900.559	261.233	75.5167	212.267	1	38	0	0	0	019SyqXtp24FW5gx
1	0	17	61.9	45.1	46.8	34.7	0	1	0	0	0	4	4	61	0	0	0	56.05	0.83333	1	16	0	0	0	01AcjebwYtXaQHmT
1	0	18	63	61.2	65.4	46.7	0	1	0	0	0	4	1	100	0	7235.54	218.567	130.367	88.2	1	36	0	0	0	01amktWPkRg2ZdYc
4	0	37	76	76	132.4	196	944.035	0	9	12	0	4	3	150	400	944.857	67.85	0	67.85	2	35	0	0	0	01C1m7ipZedqk5Ij
1	0	16	45.3	46.6	38.8	51	0	1	0	0	0	4	1	100	0	3993.49	41.2667	0	76.5	1	23	0	0	0	01eB9RmjZUQ8Vaix
1	0	6	80.7	81.5	48.7	81.8	0	1	0	0	0	4	1	50	0	2379.94	155.7	368.15	0	2	42	0	0	0	01EfjaH9AgThcxdN
4	0	102	132.2	219.15	196.7	245	4282.11	0	12	24	0	4	3	230	1916.95	1998.28	179.483	73.25	438.767	2	40	2	1	0	01EmGA6L9OckwhSY
1	0	4	44.9	34.06	29.85	0	0	1	0	0	0	4	1	20	0	0	0	43.6833	0	1	19	0	0	0	01Fd9nU2cLMGWlrp
1	0	10	60.9	64.1	51	61.7	0	1	0	0	0	4	3	90	0	10124.5	38.5667	38.5667	0	2	21	0	0	0	01fGXa9AO2PNgzVH
4	0	47	110	100	100	100	6346.85	0	12	24	0	4	1	100	140.853	6361.79	253.417	0	308.883	1	42	1	2	0	01ftmLkECBuVOIjJ
4	0	50	76	76	76	80	132.895	0	0	0	0	4	1	76	400	132.539	0.81667	0	9.08333	1	47	0	0	0	01gdwE8ZvMnCJRsi
4	0	64	96	96	96	96	0	0	3	12	0	4	3	204	0	527.815	87.6333	0	87.6333	1	37	0	0	0	01GKicdpkIQCSs3g
1	0	12	60.7	54.3	53.3	111	0	1	0	0	0	4	3	60	0	0	0	81.7167	0	1	22	0	0	0	01HNBSwsKzDtAOem
4	0	95	76	76	76	76	898.567	0	1	24	0	4	1	20	800	437.595	30.9167	0	46.0667	1	63	0	0	0	01hQibuBAYxJprTL
4	1	151	106	106	106	106	108.927	0	1	24	0	4	1	100	0	1741.47	91	0	91	1	60	0	0	0	01If4Ede6Ysz5KX9
1	0	10	73.9	29.4	20	55.2	0	1	0	0	0	4	3	100	0	0	0	179.55	0	1	25	0	0	0	01INKOnYdSQR2g9o
1	0	4	19.19	12.66	19.8	17.54	0	1	0	0	0	4	2	30	0	345.981	0.23333	17.3833	0	2	20	0	0	0	01jKhOyg8QtNVib4
1	0	17	81.5	84.7	80.5	69.2	0	1	0	0	0	4	3	100	0	54660.2	71.0167	71.0167	0	1	24	0	0	0	01K7zbGfudXnC6wN
4	0	75	136	136	136	136	0	0	0	0	0	3	2	200	1024	0	0	0	65.15	1	29	0	0	0	01KiGcOtWrzCq69L
1	0	18	73.8	73	48.1	69.5	0	1	0	0	0	4	2	100	0	9713.37	20	87.9833	0	1	22	0	0	0	01kqVnh5EeByadjD
1	0	10	75.1	93.2	82.9	81.3	0	1	0	0	0	4	1	30	0	19.2886	0	0	0	1	36	0	0	0	01LP98yBzSoGufb5
1	0	9	44.3	46.5	32.8	49.3	0	1	0	0	0	4	12	45.92	0	0	0	32.3	0	1	17	0	0	0	01NQTeJRyBKq5CPM
4	1	73	106	106.2	106	107.74	13215.2	0	1	24	1	4	2	114.81	1024	13265.3	26.1	0	26.1	2	30	0	0	0	01Q8SgVqDRJGoMlp
4	1	36	110.05	106	326.84	287.64	0	0	1	24	0	4	2	150	0	3413.23	149.467	24.5667	304.767	2	36	1	2	0	01r7OLNDhCdYP2ge
1	0	15	59.4	65.1	45.6	61.6	0	1	0	0	0	4	3	90	0	4547.36	95.3	97.55	0	2	22	0	0	0	01RrAiFTs6udXHec
1	0	11	130.8	100.6	84.9	63	0	1	0	0	0	4	2	150	0	19729.9	3.28333	91.8167	0	1	23	0	0	0	01ScqwOTh92fHnzi

图 4-1　中国联通套餐 test 数据展示

4.4 数据属性

本研究所用的资料可以从网页下载，下载资料集档案采用 CSV 档，用 Excel 将资料进行分类。该资料集共有两个数据集。一个是测试集：共包含 160567 条线。另一个是训练集：共包含 374656 条线，每一条线代表一个用户选择的套餐信息。数据集集合了各种数据类型。此数据集的数据共有 26 个类别，分别是服务类型、是不是混合服务、在线时间、当月总费用、前第 1 个月总费用、前第 2 个月总费用、前第 3 个月总费用、当月累计流量、是否连续超套、合约类型、合约时间、是否是承诺低消耗、网络口径用户、支付次数、支付金额、上个月的结转流量、本地当月流量、本地通话时长、呼叫服务时长 1、呼叫服务时长 2、性别、年龄、投诉级别、前投诉次数、以前的补救费、用户 ID。

由于中国联通套餐的数据集有两个，但是两个数据集里的属性基本都一样，所以我们只介绍测试集的属性。

（1）service_type：服务类型，为数值类型，表示联通套餐的种类，范围 1 和 4。

（2）is_mix_service：是不是固移融和套餐，为数值类型，表示联通套餐用户是不是固移融和套餐，范围 0-1。

（3）online_time：在线时间，为数值类型，表示联通套餐用户的在线时间，

范围 0-无穷。

（4）1_total_fee：当月总费用，为数值类型，表示联通套餐用户的当月总费用，范围 0-无穷。

（5）2_total_fee：前第 1 个月总费用，为数值类型，表示联通套餐用户的前第 1 个月总费用，范围 0-无穷。

（6）3_total_fee：前第 2 个月总费用，为数值类型，表示联通套餐用户的前第 2 个月总费用，范围 0-无穷 4。

（7）4_total_fee：前第 3 个月总费用，为数值类型，表示联通套餐用户的总费用 3，范围 0-无穷。

（8）month_traffic：当月累计流量，为数值类型，表示联通套餐用户的每个月的当月累计流量，范围 0-无穷。

（9）many_over_bill：连续超套，为数值类型，表示联通套餐用户是否连续超套，范围 0-1。

（10）contract_type：合约类型，为数值类型，表示联通套餐用户的合约类型，范围 0-12。

（11）contract_time：合约时间，为数值类型，表示联通套餐用户的合约时间，范围 0-24。

（12）is_promise_low_consume：是否是承诺低消耗，为数值类型，表示联通套餐用户是否是承诺低消耗，范围 0-1。

（13）net_service：网络口径用户，为数值类型，表示联通套餐用户的网络口径用户，范围 0-4。

（14）pay_times：支付次数，为数值类型，表示联通套餐用户的支付次数，范围 0-无穷。

（15）pay_num：支付金额，为数值类型，表示联通套餐用户的支付金额，范围 0-无穷。

（16）last_month_traffic：上个月的结转流量，为数值类型，表示联通套餐用户的上个月的结转流量，范围 0-无穷。

（17）local_trafffic_month：本地当月流量，为数值类型，表示联通套餐用户的本地当月流量，范围 0-无穷。

（18）local_caller_time：本地通话时长，为数值类型，表示联通套餐用户的本地通话时长，范围 0-无穷。

（19）service1_caller_time：呼叫服务时长 1，为数值类型，表示联通套餐用户的呼叫服务时长 1，范围 0-无穷。

（20）service2_caller_time：呼叫服务时长 2，为数值类型，表示联通套餐用户的呼叫服务时长 2，范围 0-无穷。

（21）gender：性别，为数值类型，表示联通套餐用户的性别，范围 1-2。

（22）age：年龄，为数值类型，表示联通套餐用户的年龄，范围 0-150。

（23）complaint_level：投诉级别，为数值类型，表示联通套餐用户的投诉级别，范围 0-2。

（24）former_complaint_num：前投诉次数，为数值类型，表示联通套餐用户的前投诉次数，范围 0-无穷（次）。

（25）former_complaint_fee：以前的补救费，为数值类型，表示联通套餐用户的以前的补救费，范围 0-无穷（分）。

（26）user_id：用户 ID，为字符类型，表示联通套餐用户的用户 ID。

而训练集比测试集就多了一个属性。

（27）current_service：当前服务，为数值类型，表示联通套餐用户的当前服务。

由于我的是中国联通套餐的数据集，最后肯定要预测分析出用户消费习惯及偏好，匹配用户最合适的套餐，提升用户感知，带动用户需求，从而帮助用户发现合适套餐，也能将合适套餐信息推送给用户。所以要对数据集进行预处理。

4.5 数据预处理

在网络下载后，我们得到的数据会存在有缺失值、重复值等，在使用之前需要进行数据预处理。数据预处理没有标准的流程，通常针对不同的任务和数据集属性的不同而不同。数据预处理的常用流程为：去除唯一属性、处理缺失值、属性编码、数据标准化正则化、特征选择、主成分分析。

数据预处理的主要步骤分为：数据清理、数据集成、数据规约和数据变换。

4.5.1 去除唯一属性

唯一属性通常是一些 id 属性，这些属性并不能刻画样本自身的分布规律，所以简单地删除这些属性即可。

4.5.2 处理缺失值

缺失值使数据记录丢失了部分信息，一些鲁棒性不佳的模型也会因为缺失值而导致无法计算数据。缺失值的处理，一般有以下两种思路：丢弃和估计。

1 丢弃

你可以只丢弃缺失项处的值，也可以丢弃包含缺失项的整条数据记录，这得看该条数据记录上其它的数据是否有价值，尤其是在数据样本较少的情况下，需要权衡一番。

2 估计

不想丢弃缺失值时，对缺失值进行估计是必要的。估计的方法有多种，最直接的是让有经验的人员手工填写，除此之外其它的常见方法有如下几种：

(1)替代。用缺失值所处属性上全部值的平均值（此时也可以加权重）、某个分位值代替。对于时间序列，则可以用相邻数据记录处值（或平均值）替代。

(2)填充。可以用与缺失值记录“相似”记录上的值来填充缺失值，不过这里需要先定义“相似”，这可能会是一个棘手的问题，用 K 最邻近、聚类等方法估计缺失值都是这种思想。对于时间序列，则可以用插值的方法，包括线性和非线性插值。

(3)基于统计模型的估计。基于非缺失的值构建统计模型，并对模型参数进行估计，然后再预测缺失处的值。

4.5.3 属性编码

（1）特征二元化

特征二元化的过程是将数值型的属性转换为布尔值的属性，设定一个阈值作为划分属性值为 0 和 1 的分隔点。

（2）独热编码（One-HotEncoding）

独热编码采用 N 位状态寄存器来对 N 个可能的取值进行编码，每个状态都由独立的寄存器来表示，并且在任意时刻只有其中一位有效。

独热编码的优点：能够处理非数值属性；在一定程度上扩充了特征；编码后的属性是稀疏的，存在大量的零元分量。

4.5.4 数据标准化正则化

（1）数据标准化：是将样本的属性缩放到某个指定的范围。

数据标准化的原因：某些算法要求样本具有零均值和单位方差；需要消除样本不同属性具有不同量级时的影响：①数量级的差异将导致量级较大的属性占据主导地位；②数量级的差异将导致迭代收敛速度减慢；③依赖于样本距离的算法对于数据的数量级非常敏感。

min-max 标准化（归一化）：对于每个属性，设 minA 和 maxA 分别为属性 A 的最小值和最大值，将 A 的一个原始值 x 通过 min-max 标准化映射成在区间[0,1]中的值 x'，其公式为：新数据=（原数据-最小值）/（最大值-最小值）。

z-score 标准化（规范化）：基于原始数据的均值（mean）和标准差（standarddeviation）进行数据的标准化。将 A 的原始值 x 使用 z-score 标准化到 x'。z-score 标准化方法适用于属性 A 的最大值和最小值未知的情况，或有超出取值范围的离群数据的情况。新数据=（原数据-均值）/标准差。

$$\mu^{(i)}=\frac{1}{N}\sum_{i=1}^{N}x_i^{(j)}$$

。

$$\sigma^{(j)}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i^{(j)}-\mu^{(j)})^2}$$

。

均值和标准差都是在样本集上定义的，而不是在单个样本上定义的。标准化是针对某个属性的，需要用到所有样本在该属性上的值。

（2）正则化：是将样本的某个范数（如 L1 范数）缩放到到位 1，正则化的过程是针对单个样本的，对于每个样本将样本缩放到单位范数。

设数据集

$$D=\{(\vec{\boldsymbol{x}}_1,y_1),(\vec{\boldsymbol{x}}_2,y_2),\cdots,(\vec{\boldsymbol{x}}_N,y_N)\},\vec{\boldsymbol{x}}_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(d)})^T$$

。对样本首先计算 Lp 范数：$L_p(\vec{\boldsymbol{x}}_i)=(\left|x_i^{(1)}\right|^p+\left|x_i^{(2)}\right|^p+\cdots+\left|x_i^{(d)}\right|^p)^{\frac{1}{p}}$。

正则化后的结果为：每个属性值除以其 Lp 范数：

$$\hat{\vec{\boldsymbol{x}}}_i=(\frac{x_i^{(1)}}{L_p(\vec{\boldsymbol{x}}_i)},\frac{x_i^{(2)}}{L_p(\vec{\boldsymbol{x}}_i)},\cdots,\frac{x_i^{(d)}}{L_p(\vec{\boldsymbol{x}}_i)})^T$$

。

4.5.5 特征选择（降维）

从给定的特征集合中选出相关特征子集的过程称为特征选择。进行特征选择的两个主要原因是：减轻维数灾难问题和降低学习任务的难度。进行特征选择必须确保不丢失重要特征。而常见的特征选择类型分为三类：过滤式（filter）、包裹式（wrapper）、嵌入式（embedding）。

过滤式选择：该方法先对数据集进行特征选择，然后再训练学习器。特征选择过程与后续学习器无关。Relief 是一种著名的过滤式特征选择方法。

包裹式选择：该方法直接把最终将要使用的学习器的性能作为特征子集的评价原则。其优点是直接针对特定学习器进行优化，因此通常包裹式特征选择比过滤式特征选择更好，缺点是由于特征选择过程需要多次训练学习器，故计算开销要比过滤式特征选择要大得多。

嵌入式选择：号称结合了过滤式和包裹式的优点，将特征选择嵌入到模型构建的过程中： 这是特征选择的一整个流程的总结，所谓嵌入式特征选择，就是通过一些特殊的模型拟合数据然后根据模型自身的某些对于特征的评价的属性来作为评价指标，最后再使用包裹式的特征选择方法来选择，当然，很多时候我们还是仅停留在计算出评价指标的阶段，因为包裹式特征选择的最大问题就是计算量和时间是三者之中最大的。

常见的降维方法：SVD、PCA、LDA。

4.5.6 主成分分析

PCA（Principal Component Analysis）是一种常用的数据分析方法。PCA 通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维。

设有 m 条 n 维数据。

1）将原始数据按列组成 n 行 m 列矩阵 X

2）将 X 的每一行（代表一个属性字段）进行零均值化，即减去这一行的均值

3）求出协方差矩阵 C=\frac{1}{m}XX^\mathsf{T}

4）求出协方差矩阵的特征值及对应的特征向量

5）将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 k 行组成矩阵 P

6）Y=PX 即为降维到 k 维后的数据。

4.6 预处理过程

我的预处理过程包含两步，第一是在电脑上安装 pycharm 和 pip，并创建项目。第二是运行预处理代码。

4.6.1 预处理工具

首先，要在电脑上安装 pycharm 工具和 annacinda 工具，这些工具直接上官网就可以直接下载。然后把创建项目，在创建项目的他是把 annacinda 添加到 pycharm 中。见图 4-2。

Pycharm 下载地址为 https://www.jetbrains.com/pycharm/。

annacinda 下载地址为 https://www.anaconda.com/。

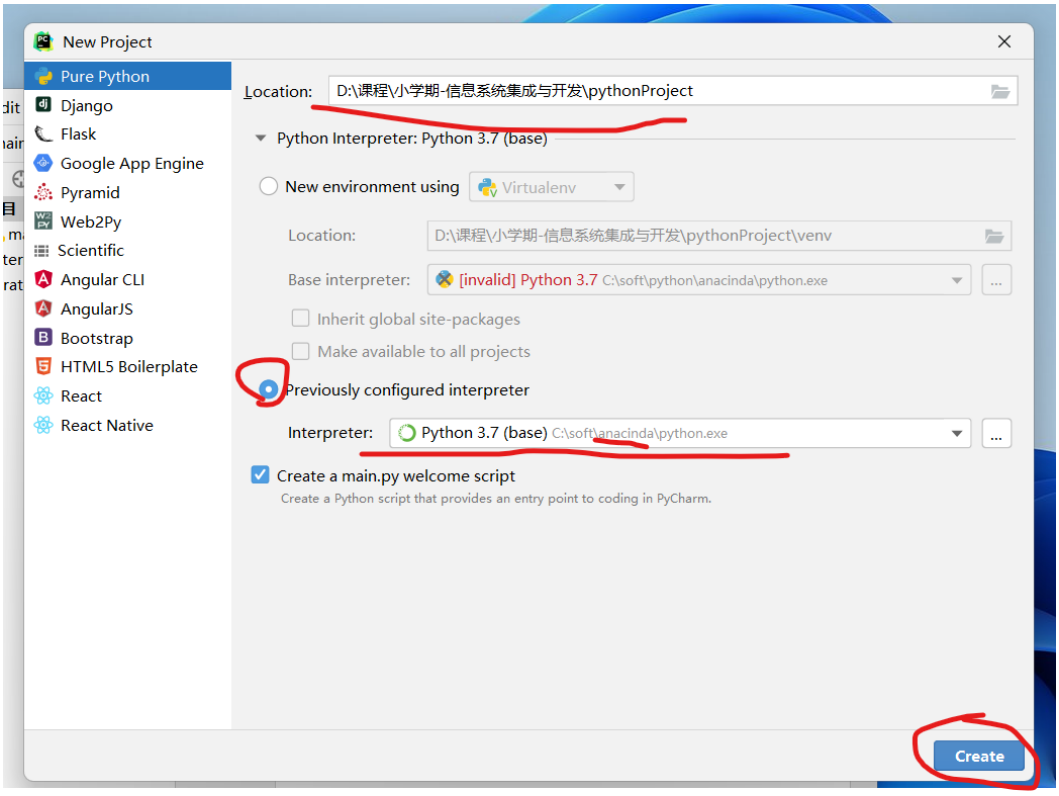


图 4-2　创建项目并导入 annacinda

4.6.2 预处理代码

(1)预处理需要的库

```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
```

（2）test.csv 预处理代码。

```
#按文件读取整个文件
data=pd.read_csv('LTtest.csv',low_memory=False)
#去除唯一属性
data=data.drop('user_id',axis=1)
#删除含空的行和列：缺失值
data=data.dropna()
com=data.columns
data=MinMaxScaler().fit_transform(data)
data=pd.DataFrame(data)
data.columns=com
data.head()
```

（3）train.csv 预处理代码。

```
#按文件读取整个文件
data=pd.read_csv('LTtrain.csv',low_memory=False)
#去除唯一属性
data=data.drop('user_id',axis=1)
data=data.drop('current_service',axis=1)
#删除含空的行和列：缺失值
data=data.dropna()
com=data.columns
data=MinMaxScaler().fit_transform(data)
data=pd.DataFrame(data)
data.columns=com
data.head()
```

5　中国联通套餐的分析和预测及可视化

本小节主要介绍了如何将中国联通套餐的数据集转换为 arff 格式,算法模型评估，随机森林模型介绍，贝叶斯模型介绍，两种模型的对比，以及根据这个联通套餐数据集最后得出的结论。

5.1　中国联通套餐的数据集转换为 arff 格式

打开 Weka，选择 explorer 模式，这时会打开新的 explorer 窗口，我们点击左上角 Open File 选项，在 Look In 中找到你存放 test.csv 的位置，并且把底下的 File of Type 改成 CSV data files 格式，找到文件并打开。

最后，点击右上角的 Save…按钮，把 File Name 中的 test.csv 的后缀改成 arff，然后再点击保存，这样我们的数据就保存为了 arff 格式。

train.arff 训练集的格式同理可得。

5.2　算法模型评估

机器学习(NLP)等 AI 领域，评估(evaluation)是一项非常重要的工作，其模型或算法的评价指标往往有如下几点：准确率(Accuracy)，精确率(Precision)，召回率(Recall)和综合评价指标(F1-Measure)。

5.2.1　准确率（Accuracy）

准确率(Accuracy)是一个用于评估分类模型的指标。大白话来说就是，模型预测正确数量所占总量的比例。

准确率的伪公式：在二元分类中，可根据正类别与负类别按如下方式计算：

注：下方公式中，TP =真正例，TN =真负例，FP =假正例，FN =假负例。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

还有，当我们使用分类不平衡的数据集（如：正类别标签与负类别标签数量存在明显差异）时，就一项准确率并不能反映情况。

为了更好的评估分类不平衡的数据集问题，下面引入精确率（Precision）和召回率（Recall）

5.2.2 精确率（Precision）

精确率为解决在被识别为正类别的样本中，为正类别的比例。精确率的公式定义如下：注：下方公式中，TP =真正例，FP =假正例。

$$\text{Precision} = \frac{TP}{TP + FP}$$

注：如果模型预测结果中没有假正例，则模型的精确率为 1。

5.5.3 召回率（Recall）

召回率为解决在所有正类别样本中，被正确识别为正类别的比例。召回率的公式定义如下：注：下方公式中，TP =真正例，FP =假正例。

$$\text{召回率} = \frac{TP}{TP + FN}$$

注：如果模型的预测结果没有假负例，则模型的召回率为 1。

想要全面评估模型的有效性，必须同时检查精确率与召回率。但是，很遗憾，精确率和召回率往往是此消彼长。也就是说，提高精确率通常会降低召回率，反之亦然。

5.5.4 综合评价指标（F1-Measure）

F-Measure 是一种统计量，又称 F-Score,也是精确率（Presicion）和召回率（Recall）的加权调和平均，常用于评价分类模型的好坏。

F-Measure 数学公式为：注：下方公式中，P 为 Precision，R 为 Recall，a 为权重因子。

$$F = \frac{(a^2 + 1) * P * R}{a^2 * (P + R)}$$

当 a=1 时，F 值变为最常见的 F1 了，代表精确率和召回率的权重一样，是最常见的一种评价指标，因此，F1 的数学公式为：

$$F1 = \frac{2 * P * R}{P + R}$$

F1 综合了精确率和召回率的结果，当 F1 较高时，则说明模型或算法的效果比较理想。

5.3 随机森林模型

RandomForest 应该算是一个特别简单但是有效的算法，其核心思想是通过训练和组合不同的决策树，形成森林，最后的分类结果由这多棵树进行投票来决定。

（1）训练：首先，对训练集进行有放回的抽样 N 次，得到训练集的一个子集作为新训练集。

然后，在新的训练集中随机抽出训练集的 K 个属性，训练一棵分类树，并且不对这个分类树做剪枝操作。

最后，重复上述过程 M 次，得到 M 个分类器。

（2）判定：对于任意一个新的用例，使用 M 个分类器进行分类，最后的分类结果由这 M 个分类器投票决定。

可以看出，RandomForest 在 Bagging 的基础上，主要是增加了随机抽出 K 个属性进行训练，从经验上来讲，假设属性总量为 X，则要求 K<<X，一般取 K=sqrt(X)。

（3）经过这样的改进，RandomForest 又有了如下的一些优点：

第一，可以处理高纬度的数据，显而易见，每次抽取 K 个属性进行训练，提高训练速度。

第二，可以评估每个属性的重要程度，根据子分类树的精确度，就可以评估属性的重要程度。

第三，对于属性的遗失，可以很好的处理，因为各子分类树构建在不同的属性之上，可以只挑选一部分可用的子分类树进行判定

5.4 贝叶斯模型

贝叶斯是机器学习中的一个重要分支。经过多年的研究，该方法已被广泛地应用于各个领域，特别是在专家系统方面，获得了很好的效果。由于其在理论上能最大限度地减少错误，因而在分类问题中得到了越来越多的应用，因而又一次成为了数据挖掘的热门话题。

贝叶斯推理是以概率为基础的一种推理方式，其理论依据是与概率有关的理论。有条件概率是一种在日常生活中经常使用的方法。比如，既然知道了今天要下雨，那明天还有什么可能？这个问题涉及到了有条件的可能性。一般情况下，如果 A 事件已发生，则 B 的可能性有多大？我们把它叫做 B 事件的有条件的可能性。其定义如下：假定 A、B 为两个事件，且知道 A 发生的可能性 P(A)>0，则 B 事件的发生几率为：

$$P(B|A) = \frac{P(AB)}{P(A)}$$

在某些情况下，我们必须考虑两个事件同时出现的可能性，也就是 A 与 B 的合并几率是 P（AB）。它的公式为：

$$P(AB) = P(A)P(B|A)$$

也就是当 A 事件发生时，B 事件也发生的概率，即两个事件同时发生的概率。假设影响事件 A 的事件有$B_1, B_2 \ldots B_n$并且满足下面的条件：$B_1 \cap B_2 = \emptyset$，$P \cup B_1 = 1$，$P(B_i) > 0$，$i > 1,2,3 \ldots n,$则有如下全概率公式：

$$P(A) = \sum_{i=0}^{n} n\, P(B_i)P(A|B_i)$$

虽然我们并不清楚其中的一些可能性，但我们可以通过我们所掌握的资料和自己的经历来估算每个事件发生的机率。这个可能性叫做“先验”。若从已有资料中得到的先验机率，则此机率即为先验机率；如果用我们的经验来判断，这个可能性被称为一个先验的可能性。

后验概率是指在调查中利用贝叶斯公式修正前验概率，最后获得更为精确的概率，也就是后验概率。

5.5 模型对比

下面将使用随机森林算法模型和贝叶斯模型进行对比，以得出每个模型的优势与不足，完成联通行业用户套餐的预测与推荐任务。由于中国联通套餐的数据集有两个，所以我采用第二种测试方法，一部分用来学习，一部分用来考试。

5.5.1 随机森林模型

随机森林模型的准确率(Accuracy)，精确率(Precision)，召回率(Recall)和综合评价指标(F1-Measure)见图 5-1。其套餐类型为 1 的准确率为 0.665，精确率为 0.911，由于模型的预测结果有假负例，所以随机森林模型的召回率为 0.950，而随机森林模型的综合评价指标为 0.930。其套餐类型为 4 的准确率为 0.665，精确率为 0893，由于模型的预测结果有假负例，所以随机森林模型的召回率为 0.816，而随机森林模型的综合评价指标为 0.853。

```
=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.950    0.184    0.911      0.950    0.930      0.785    0.965     0.981     1
                0.816    0.050    0.893      0.816    0.853      0.785    0.965     0.934     4
Weighted Avg.   0.905    0.139    0.905      0.905    0.904      0.785    0.965     0.965
```

图 5-1　中国联通套餐的随机森林的算法模型评估

随机森林模型的混淆矩阵见图 5-2。248959 表示是在训练集中原本有 248959 个 1，被正确的分类为 1，0 表示是在训练集中原本有 0 个 1，被错误的分类为 4，125659 表示是在训练集中原本有 125659 个 4，被错误的分类为 1，0 表示是在训练集中原本有 0 个 4，被正确的分类为 4。

```
=== Confusion Matrix ===

      a      b   <-- classified as
 236633  12326 |      a = 1
  23142 102517 |      b = 4
```

图 5-2　中国联通套餐的随机森林的混淆矩阵

随机森林模型的可视化图见图 5-3。蓝色是套餐类型为 1 的，红色是套餐类型为 4 的，由可视化图可以看出来基本上都是蓝色比红色多。

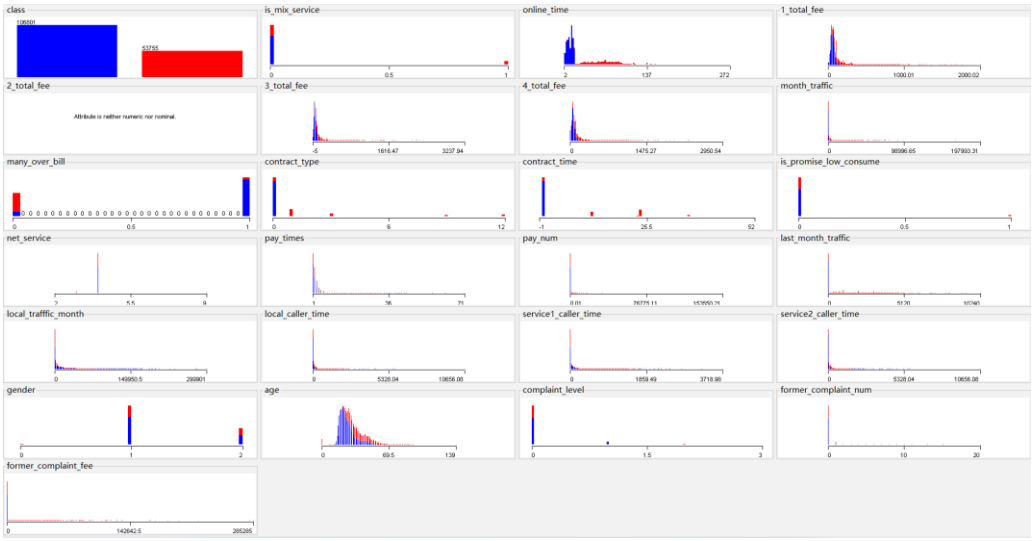


图 5-3　随机森林模型的可视化图

5.5.2　贝叶斯模型

贝叶斯模型的准确率(Accuracy)，精确率(Precision)，召回率(Recall)和综合评价指标(F1-Measure)见图 5-4。其套餐类型为 1 的准确率为 0.664，精确率为 0.882，由于模型的预测结果有假负例，所以贝叶斯模型的召回率为 0.943，而贝叶斯模型的综合评价指标为 0.911。其套餐类型为 4 的准确率为 0.664，精确率为 0869，由于模型的预测结果有假负例，所以随机森林模型的召回率为 0.750，而随机森林模型的综合评价指标为 0.805。

```
=== Detailed Accuracy By Class ===

                TP Rate   FP Rate   Precision   Recall   F-Measure   MCC      ROC Area   PRC Area   Class
                0.943     0.250     0.882       0.943    0.911       0.721    0.937      0.963      1
                0.750     0.057     0.869       0.750    0.805       0.721    0.937      0.899      4
Weighted Avg.   0.878     0.185     0.878       0.878    0.876       0.721    0.937      0.941
```

图 5-4　中国联通套餐的贝叶斯的算法模型评估

贝叶斯模型的混淆矩阵见图 5-5。234765 表示是在训练集中原本有 234765 个 1，被正确的分类为 1，14194 表示是在训练集中原本有 14194 个 1，被错误的分类为 4，31451 表示是在训练集中原本有 31451 个 4，被错误的分类为 1，94208 表示是在训练集中原本有 94208 个 4，被正确的分类为 4。

```
=== Confusion Matrix ===

      a      b   <-- classified as
 234765  14194 |      a = 1
  31451  94208 |      b = 4
```

图 5-5　中国联通套餐的贝叶斯的混淆矩阵

贝叶斯模型的可视化图见图 5-6。蓝色是套餐类型为 1 的，红色是套餐类型为 4 的，由可视化图可以看出来基本上都是蓝色多余红色。

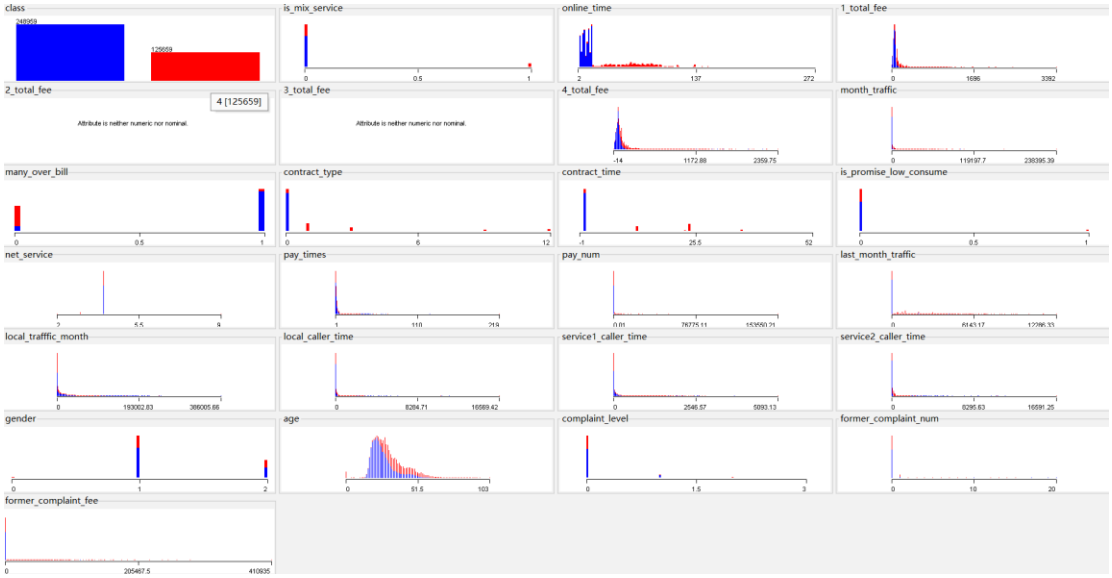


图 5-6　贝叶斯模型的可视化图

5.5.3　对比

由上述两种模型的的准确率(Accuracy)，精确率(Precision)，召回率(Recall)和综合评价指标(F1-Measure)对比，混淆矩阵对比以及可视化图的对比可以看出随机

森林模型比贝叶斯模型要好一点，而且还可以看出不管是算法模型评估的四个指标，还是混淆矩阵和可视化图套餐类型为 1 的要比套餐类型为 4 的要受欢迎的多。

5.6 结论

根据 Weka 工具我使用随机森林算法模型和贝叶斯模型进行对比，从而得出了的结论为：

从随机森林模型的准确率(Accuracy)，精确率(Precision)，召回率(Recall)和综合评价指标(F1-Measure)，混淆矩阵和可视化图以及贝叶斯模型的准确率(Accuracy)，精确率(Precision)，召回率(Recall)和综合评价指标(F1-Measure)，混淆矩阵和可视化图可以预测分析出用户消费习惯及偏好更偏向于套餐类型为 1 的，而匹配到的用户最合适的套餐也是套餐类型为 1 的。所以为了提升用户感知，带动用户需求，要将套餐类型为 1 的的套餐信息推送给用户或把套餐类型为 1 的套餐信息放在用户可以明显找到的地方。

参考文献

[1] 百度百科, URL: https://baike.baidu.com/item/Weka/1070121. [2022-6-16].

[2] 博客, URL: https://blog.csdn.net/shaoz/article/details/6841498?. [2022-6-16].

[3] 博客, URL: https://blog.csdn.net/sinat_25873421/article/details/82724903. [2022-6-17].

[4] 简书, URL: https://www.jianshu.com/p/73c6fce1dbe9. [2022-6-16].

[5] 百度百科, URL: https://baike.baidu.com/item/DataFountain/58361468. [2022-6-21].

[6] 知乎, URL: https://zhuanlan.zhihu.com/p/51782481. [2022-6-20].

[7] 博客, URL: https://blog.csdn.net/binbigdata/article/details/84565486. [2022-6-22].

[8] 博客, URL: https://blog.csdn.net/huguozhiengr/article/details/85162268. [2022-6-22].

[9] 博客, URL: https://blog.csdn.net/u014772862/article/details/52335970 [2022-6-22].

[10] B 战, URL: https://www.bilibili.com/video/BV1ki4y1C71H. [2022-6-23].

[11] 博客, URL: https://blog.csdn.net/qq_17753903/article/details/89817371. [2022-6-23].

[12] 知乎, URL: https://zhuanlan.zhihu.com/p/62180318. [2022-6-23].

[13] 博客, URL: https://blog.csdn.net/smilehehe110/article/details/54425787. [2022-6-23].

[14] 知乎, URL: https://zhuanlan.zhihu.com/p/37575364. [2022-6-23].

[15] 简书, URL: https://www.jianshu.com/p/d6d5e7d423e7. [2022-6-23].