数据挖掘作业二：关联规则挖掘

贺鹏飞　　3220180700

1.数据源

选择 Wine Reviews（https://www.kaggle.com/zynicide/wine-reviews）作为关联规则挖掘的数据源，该数据源包含超过 15 万条记录，每条记录包含 country、point、price、province、region、variety、winery 等属性。首先将数据源导入到 matlab 中，并转换成方便处理的格式。相关代码：WineReviewsImport.m

```
%% 导入文本文件中的数据。
% 用于从以下文本文件导入数据的脚本:
%
%       E:\MyWorkplace\DataMining\wine-reviews\winemag-data_first150k.csv
%
%% 初始化变量。
filename = 'E:\MyWorkplace\DataMining\wine-reviews\winemag-data_first150k.csv';
delimiter = ',';
startRow = 2;

%% 每个文本行的格式:
%     列 1: 双精度值 (%f)
%     列 2: 文本 (%q)
%     列 3: 文本 (%q)
%     列 4: 文本 (%q)
%     列 5: 双精度值 (%f)
%     列 6: 双精度值 (%f)
%     列 7: 文本 (%q)
%     列 8: 文本 (%q)
%     列 9: 文本 (%q)
%     列 10: 文本 (%q)
%     列 11: 文本 (%q)
formatSpec = '%f%q%q%q%f%f%q%q%q%q%q%[^\n\r]';

%% 打开文本文件。
fileID = fopen(filename,'r');

%% 根据格式读取数据列。
dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter, 'TextType', 'string', 'EmptyValue', NaN, 'HeaderLines' ,startRow-1, 'ReturnOnError', false, 'EndOfLine', '\r\n');

%% 关闭文本文件。
fclose(fileID);
```

```
%% 将导入的数组分配给列变量名称
VarName1 = dataArray{:, 1};
country = cellstr(dataArray{:, 2});
description = cellstr(dataArray{:, 3});
designation = cellstr(dataArray{:, 4});
points = dataArray{:, 5};
price = dataArray{:, 6};
province = cellstr(dataArray{:, 7});
region_1 = cellstr(dataArray{:, 8});
region_2 = cellstr(dataArray{:, 9});
variety = cellstr(dataArray{:, 10});
winery = cellstr(dataArray{:, 11});

%% 清除临时变量
clearvars filename delimiter startRow formatSpec fileID dataArray ans;
```

2.数据集预处理

对数据集进行处理，转换成适合关联规则挖掘的形式。

由于数值属性 points 和 price 分布过于分散，不适合关联规则挖掘，首先对这两个属性进行处理，通过对 points 属性的统计，可以发现 points 的取值范围是[80,100],并且接近正态分布，因此将 points 分为 5 个级别，如表 1 所示：

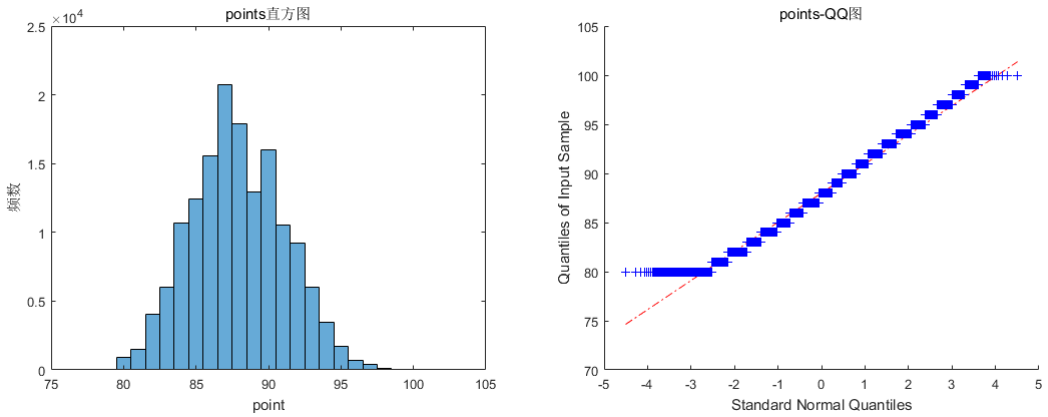

图 1.points 属性的直方图和 QQ 图

表 1.points 属性的预处理

Points	[80,84]	[85,89]	[90,94]	[95,99]	100
级别	a	b	c	d	e

同理，观察 price 的直方图分布，可以发现 price 的取值大部分集中在[4,100]的区间内，因此，将 price 划分为 5 个不等间隔的级别。

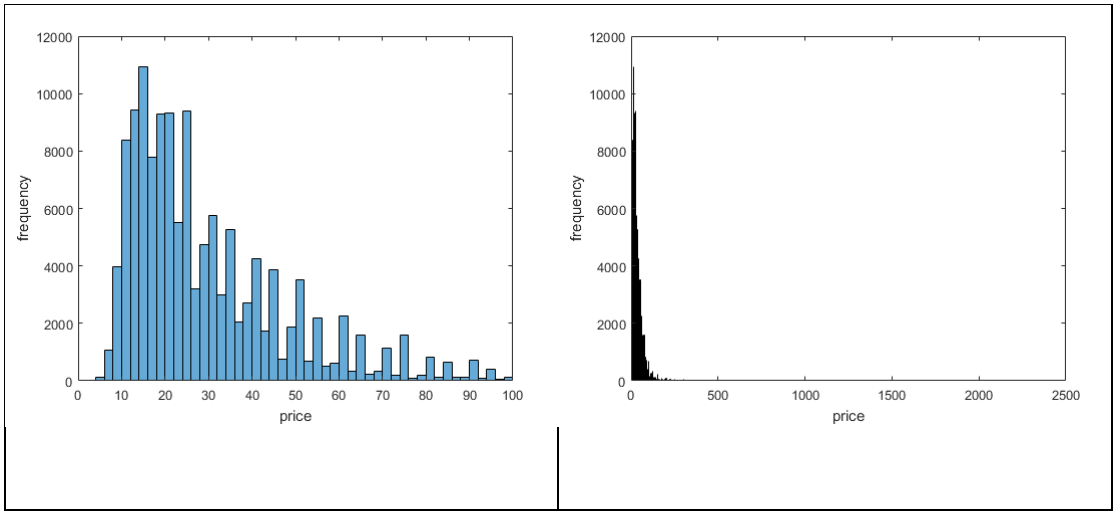

图 2.price 属性的直方图

表 2.price 属性的预处理

Price	[0,50]	(50,100]	(100,500]	(500,1000]	>1000
级别	A	B	C	D	E

DataTrans.m

```
%% 将 points 和 price 划分为 5 个级别
[n,~]=size(price);
price_level=cell(n,1);
points_level=cell(n,1);
points_level(points>=80&points<85)={'a'};
points_level(points>=85&points<90)={'b'};
points_level(points>=90&points<95)={'c'};
points_level(points>=95&points<100)={'d'};
points_level(points==100)={'e'};
price_level(price<=50)={'A'};
price_level((price>50)&(price<=100))={'B'};
price_level(price>100&price<=500)={'C'};
price_level(price>500&price<=1000)={'D'};
price_level(price>1000)={'E'};
%%将处理后的数据保存到文件 data.txt
fulldata=[country,points_level,price_level,province,variety,winery];
data=fulldata(~(ismissing(price)|ismissing(province)),:);
[nrows,ncols]= size(data);
filename = 'data.txt';
fid = fopen(filename, 'w');
for row=1:100000
    for col=1:ncols
        if(col<ncols)
            fprintf(fid,'%s%c',data{row,col},',');
        else if (col==ncols)
```

```
                fprintf(fid,'%s\r\n',data{row,col});
            end
        end
    end
end
fclose(fid);
```

3.频繁项集和关联规则

使用 Apriori 算法寻找频繁项集并进行关联规则挖掘。Apriori 算法是关联规则最常用也是最经典的挖掘频繁项集的算法，其核心思想是通过连接产生候选项及其支持度，然后通过剪枝生成频繁项集。

Apriori 算法的实现分为两个过程：1.找出所有的频繁项集（支持度必须大于等于给定的最小支持度阈值），在这个过程中连接步和剪枝步互相融合，最终得到最大的频繁项集。2. 由频繁项集产生强关联规则：在过程 1 可知未超过预定的最小支持度阈值的项集已被剔除，如果剩下这些规则又满足了预定的最小置信度阈值，那么就挖掘出了强关联规则。

设置最小支持度为 0.2，最小置信度为 0.3，寻找到频繁项集如表 3 所示。共生成 13 条关联规则，如表 4 所示。

主程序：cal_apriori.m，辅助程序：trans2matrix.m、findRules.m。

表 3.频繁项集

1-频繁项集	2-频繁项集	3-频繁项集
{A}	{A,California }	{A,California,US}
{California}	{A,US}	{A,US,b}
{US}	{A,b}	
{b}	{California,US}	
{c}	{US,b}	

表 4.关联规则及支持度和置信度

关联规则	支持度	置信度
A -> b	49.7624%	57.9639%
b -> A	49.7624%	94.0996%
US -> A	39.1467%	86.453%
California -> US	32.3224%	100%
US -> California	32.3224%	71.3819%
California -> A	27.1828%	84.0991%
California -> A,US	27.1828%	84.0991%
US -> A,California	27.1828%	60.0315%
A,California -> US	27.1828%	100%
A,US -> California	27.1828%	69.4384%
California,US -> A	27.1828%	84.0991%
A,US -> b	20.3862%	52.0765%
US,b -> A	20.3862%	93.1881%

关联规则评价

使用提升度 lift、全自信度和最大自信度作为评价关联规则的指标，对应结果如表 5 所示。以关联规则 A->b 为例，该规则表示价格在 50 元以下的酒，评分在 85 到 89 之间。该规则的提升度大于 1，说明是价格和评分是正相关的，全自信度和最大自信度也同样显示正相关。通过观察可以发现，价格在 50 元以下的酒，评分大都在 85 到 89 之间，而这些酒很可能产自美国的 California。

表 5.关联规则各项指标

规则	支持度	置信度	提升度	全自信度	最大自信度
A -> b	49%	57%	1.09	57%	94%
b -> A	49%	94%	1.09	57%	94%
US -> A	39%	86%	1	86%	86%
California -> US	32%	100%	2.22	71%	100%
US -> California	32%	71%	2.22	71%	100%
California -> A	27%	84%	0.97	84%	93%
California -> A,US	27%	84%	2.14	69%	84%
US -> A,California	27%	60%	2.22	60%	100%
A,California -> US	27%	100%	2.22	60%	100%
A,US -> California	27%	69%	2.14	69%	84%
California,US -> A	27%	84%	0.97	30%	84%
A,US -> b	20%	52%	0.98	39%	52%
US,b -> A	20%	93%	1.08	89%	93%

4.挖掘结果展示