数据挖掘导论
第一次作业

梁军

郑州大学

May 7, 2014

1 Wine Quality数据集

酒质量(Wine Quality)数据集是由葡萄牙米尼奥大学(Univ. Minho)P. Cortez，和CVRVV(一直致力于提高葡萄牙青酒品质的组织)的A. Cerdeira, F. Almeida, T. Matos and J. Reis 于2009年创建的[1]。该数据集中所涉及的是葡萄牙青酒——一种产自葡萄牙西北地区米尼奥的独特品种。这种酒占葡萄牙酒业产量的15%，其中约10%用于出口。该数据集采集的数据是葡萄牙青酒中两种常见的品种：白葡萄酒和红葡萄酒(按照区域划分)。数据是由CCRVV 在2004年5月到2007年2月从受保护的原产地采集测试样品并测量获取的。这些数据是通过一个叫做iLab的自动化系统记录的，它会自动管理的葡萄酒样品的测试生产要求和实验室测量、葡萄酒口感分析的过程。每个葡萄酒的口感是由三个品酒师一起评价，然后给出一个0-10的评分(0表示最坏，10表示最好)，然后从三个打分中选取一个中间值作为该葡萄酒的质量得分。

1.1 数据描述

该数据集包含1599个红葡萄酒的信息，4898个白葡萄酒的信息。每种酒的特征用下面11中属性描述：

1. fixed acidity(非挥发性酸含量)，单位：$(g(tartaric\ \ acid)/dm^3)$
2. volatile acidity(挥发性酸含量)，单位：$(g(tartaric\ \ acid)/dm^3)$
3. citric acid(柠檬酸含量)，单位：(g/dm^3)
4. residual sugar(香槟酒甜度)，单位：(g/dm^3)
5. chlorides(氯化物含量)，单位：$(g(tartaric\ \ acid)/dm^3)$
6. free sulfur dioxide(游离二氧化硫含量)，单位：(mg/dm^3)
7. total sulfur dioxide(二氧化硫总含量)，单位：(mg/dm^3)
8. density(密度)，单位：(g/cm^3)
9. pH(pH值)
10. sulphates(硫酸盐含量)，单位：$(g(potassium\ \ sulphate)/dm^3)$
11. alcohol(酒精度)，单位：$(\%vol.)$

属性	红葡萄酒			白葡萄酒		
	最小值	最大值	平均值	最小值	最大值	平均值
fixed acidity	4.6	15.9	8.3	3.8	14.2	6.9
volatile acidity	0.1	1.6	0.5	0.1	1.1	0.3
citric acid	0	1	0.3	0	1.7	0.3
residual sugar	0.9	15.5	2.5	0.6	65.8	6.4
chlorides	0.01	0.61	0.08	0.01	0.35	0.05
free sulfur dioxide	1	72	14	2	289	35
total sulfur dioxide	6	289	46	9	440	138
density	0.99	1.004	0.996	0.987	1.039	0.994
pH	2.7	4	3.3	2.7	3.8	3.1
sulphate	0.3	2	0.7	0.2	1.1	0.5
alcohol	8.4	14.9	10.4	8	14.2	10.4

1.2 属性类型

属性是对象的性质或特性，它因对象而已，或随时间而变化。属性的类型有以下四种[2]:

1. 标称：仅仅表示不同的名字；
2. 序数：提供足够的信息确定队形的序；
3. 区间：对于区间属性，值之间的差是有意义的，即存在测量单位；
4. 比率：对于比率变量，差和比率都是有意义的。

根据以上定义，我们可以将用于鉴定酒质量的11个属性分别划到对应的属性类型：

1. 比率型数据：大多数属性都属于此类型数据，如：volatile acidity(挥发性酸含量)，citric acid(柠檬酸含量)，residual sugar(香槟酒甜度)，chlorides(氯化物含量)，free sulfur dioxide(游离二氧化硫含量)，total sulfur dioxide(二氧化硫总含量)，density(密度)，sulphates(硫酸盐含量)和alcohol(酒精度)。因为对这些属性进行比率运算是有意义的，所以它们都属于比率型数据。

2. 序数型数据：酒的品质应该属于序数型数据，虽然用1-10来表示酒的品质好坏，单这仅仅是定义了一个等级，并没有测量单位，进行加法和减法运算是没有意义的，因此属于序数型数据。
3. 区间型数据?：最初感觉pH值应该属于区间型数据，和摄氏温度类似，都是定义了一个基准：摄氏温度是将冰点定为0° C，而pH值是将中性溶液指通常情况下（25° C、298K左右），pH值为7.0的溶液，常见的有氯化钠溶液、纯水定为标准溶液。但两者也有不同，比如摄氏温度其实是定义了0° C和100° C，然后将这个区间的值等分100份，这跟区间型的定义不谋而合，而pH值却不是这样，仅仅是定义了一个基准而已，因此，最终将pH 值也划归为比率型数据。

1.3 数据集特性

根据[2]中描述，我们对数据集的三个特性：维度、稀疏性和分辨率进行讨论。首先，我们来看数据集的维度，数据集的维度是数据集中的对象具有的属性数目。而对于本数据集来说，每个数据对象具有的属性数目是11，也就是该数据集的维度是11，属于低维数据，因而不存在维灾难的问题，不需要进行维规约，但是由于仅仅只有11 维可能会造成数据间的区别度不够；其次，我们来看数据的稀疏性问题，从数据集上我们可以清楚的看到，该数据集中几乎没有属性值为0的情况，也就是说该数据集不是稀疏的，对于某些仅适合处理稀疏数据的数据挖掘算法可能不适用；最后，我们看看数据的分辨率，由于数据的详细说明中并没有给出数据收集的间隔，仅告诉我们该数据集是从04年5月份到07年2月份采集的，由此可以看到该数据集的跨度是非常大的，而对象仅仅有几千个，这可能会造成数据的分辨率太低，不能有效识别出数据中的模式。

1.4 数据集类型

根据[2]中的介绍，我们知道数据集的类型可以分为以下几种：

1. 记录数据
   (a) 事务数据和购物篮数据

(b) 数据矩阵
(c) 稀疏数据矩阵

2. 基于图形的数据
   (a) 带有对象之间联系的数据
   (b) 带有图形对象的数据
3. 有序数据
   (a) 时序数据
   (b) 序数数据
   (c) 时间序列数据
   (d) 空间数据

根据数据集的特点，我们可以很容易判断出Wine Quality数据集是属于记录数据中的数据矩阵，它是一种标准的数据格式，可以使用标准的矩阵操作对数据进行变换和处理。

2 University数据集

2.1 数据描述

University数据集是Lebowitz M. 的一篇发表于机器学习上的论文中所使用的数据集[3]。是关于大学的数据集，共有285个数据对象，部分数据对象的某些属性值有缺失，对象的属性描述如下：

1. University-name(大学名称)
2. State(学校所在地)
3. location(城市规模)
4. Control(学校性质，如：私立)
5. number-of-students(学生数量)
6. male:female (ratio)(男女比例)
7. student:faculty (ratio)(学生与教职工比例)
8. sat-verbal(sat英语成绩)
9. sat-math(sat数学成绩)
10. expenses(费用)
11. percent-financial-aid(助学金比例)
12. number-of-applicants(申请人数)
13. percent-admittance(通过率)
14. percent-enrolled(入学率)

15. academics(学术规模)
16. social(社会规模)
17. quality-of-life(生活质量)
18. academic-emphasis(重点学科)

	Harvard	MIT
state	massachusetts	massachusetts
location	urban	urban
control	private	private
no-of-students thous	5-10	5-
male:female ratio	65:35	75:25
student:faculty ratio	10:1	5:1
sat verbal	700	650
sat math	675	750
expenses thous$	10+	10+
percent-financial-aid	60	50
no-applicants thous	13-17	4-7
percent-admittance	20	30
percent-enrolled	80	60
academics scale:1-5	5	5
social scale:1-5	3	3
quality-of-life scale:1-5	4	3
academic-emphasis	history	sciences
academic-emphasis	biology	electrical-engineering
academic-emphasis	liberal-arts	mechanical-engineering
academic-emphasis		engineering

2.2 属性类型

参照1.2对属性类型的定义，将该数据集的属性划分到对应的属性类型：

1. 标称型数据 此数据集中的标称型数据有：University-name、State、Control和academic-emphasis，这些属性仅仅是标识不同的名字，没有其他的信息来确定对象的序，因此属于标称型数据。
2. 序数型数据 此数据集中的序数型数据有：location、academics、social和quality-of-life，这些属性的值可以供我们确定对象的序，但由于这些属性都没有测量单位，不能进行加法和减法运算，因此数据序数型数据。
3. 比率型数据 数据集中的number-of-students、sat-verbal、sat-math、expenses和number-of-applicants都可以进行差和比率预算，因此属于比率型数据；对于数据集中的male:female、student:faculty、percent-financial-aid、percent-admittance和percent-enrolled也可以进行比率运算，但由于只是百分比，自身没有测量单位，对是否属于比率型数据有点疑问？

2.3 数据集特性

根据[2]中描述，我们对数据集的三个特性：维度、稀疏性和分辨率进行讨论。首先，我们来看数据集的维度，数据集的维度是数据集中的对象具有的属性数目。而对于本数据集来说，每个数据对象具有的属性数目是18，也就是该数据集的维度是18，属于低维数据，因而不存在维灾难的问题，不需要进行维规约，但是由于仅仅只有18 维可能会造成数据间的区别度不够；其次，我们来看数据的稀疏性问题，从数据集上我们可以看到该数据集中也几乎没有属性值为0的情况(极少对象的属性值缺失)，也就是说该数据集不是稀疏的，对于某些仅适合处理稀疏数据的数据挖掘算法可能不适用；最后，我们来看数据集的分辨率问题，美国大学数量大概有3000多所，本数据集中收集了近300所大学的信息，可以说已经收集了足够多的数据信息。

2.4 数据集类型

参照1.4对数据集类型的定义，该数据集属于记录型数据，每个大学对象相当于一条记录，每个记录包含固定的数据字段(属性)及。记录之间或数据字段之间没有明显的联系，并且每个记录(对象)具有相同的属性集。

参考文献

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4): 547 - 553, 2009.

[2] 范明, 范宏建, et al. 数据挖掘导论, 2006.

[3] Michael Lebowitz. Concept learning in a rich input domain: Generalization-based memory. Machine Learning, 1984.