数据挖掘作业

数据集选择： Wine Reviews ， winemag-data_first150k.csv 文件包含八个标称属性，分别是 'country'，'description'，'designation'，'province'，'region_1'，'region_2'， 'variety'，'winery'； winemag-data-130k-v2.csv 文件包含十一个标称属性， 分别是'country'，'description' ， 'designation' ， 'province' ， 'region_1' ， 'region_2' ， 'taster_name' ，'taster_twitter_handle'，'title'，'variety'，'winery'。

数据分析过程：

1： 数据预处理

空值处理：数量少的直接删除，多的用最大频率替换

2： 找频繁项集

对多个属性进行关联规则 挖掘，需要将来自于不同属性的值转化为可生成频繁项集的形式。采用 Apriori 算法构建频繁项集。在此任 务中，频繁项集是指经常出现在一起的属性项的集合，而一个项集的支持度 （support）定义为数据集中包含该项集的记录所占的比例。首先，规定最小 支持度（min-support）为 0.25，最小置信度（min-confidence）为 0.5。

```
def apriori(self, dataset):
    C1 = self.create_C1(dataset)
    dataset = [set(data) for data in dataset]
    L1, support_data = self.scan_D(dataset, C1)
    L = [L1]
    k = 2
    while len(L[k-2]) > 0:
        Ck = self.apriori_gen(L[k-2], k)
        Lk, support_k = self.scan_D(dataset, Ck)
        support_data.update(support_k)
        L.append(Lk)
        k += 1
    return L, support_data
```

3：计算支持度，置信度，Lift 指标

首先从一个频繁项集开始，创建一个规则列表，其中规则右部只包含一个元 素，然后对这些规则计算是否满足最小置信度要求。接下来合并所有的剩余 规则列表来创建一个新的规则列表，其中规则右部包含两个元素。最后，对 于产生的每个规则，我们分别计算其支持度(support)、置信度(confidence)以 及提升度(Lift)指标。

```python
def cal_conf(self, freq_set, H, support_data, big_rules_list):
    # 评估生成的规则
    prunedH = []
    for conseq in H:
        sup = support_data[freq_set]
        conf = sup / support_data[freq_set - conseq]
        lift = conf / support_data[freq_set - conseq]
        if conf >= min_confidence:
            big_rules_list.append((freq_set-conseq, conseq, sup, conf, lift))
            prunedH.append(conseq)
    return prunedH
```

4：可视化并分析

{"X_set": [["province", "Washington"]], "Y_set": [["country", "US"]], "sup": 0.2613656544130705, "conf": 1.0, "lift": 3.826057414642412}
{"X_set": [["province", "California"]], "Y_set": [["country", "US"]], "sup": 0.5764961818504706, "conf": 1.0, "lift": 1.7346168656141703}
{"X_set": [["taster_name", "Virginie Boone"]], "Y_set": [["country", "US"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_name", "Virginie Boone"]], "Y_set": [["province", "California"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_twitter_handle", "@vboone"]], "Y_set": [["country", "US"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_twitter_handle", "@vboone"]], "Y_set": [["province", "California"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_twitter_handle", "@vboone"]], "Y_set": [["taster_name", "Virginie Boone"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_name", "Virginie Boone"]], "Y_set": [["taster_twitter_handle", "@vboone"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_name", "Paul Gregutt"]], "Y_set": [["country", "US"]], "sup": 0.2667820990942994, "conf": 1.0, "lift": 3.7483774338492264}
{"X_set": [["taster_twitter_handle", "@paulgwine "]], "Y_set": [["country", "US"]], "sup": 0.2667820990942994, "conf": 1.0, "lift": 3.7483774338492264}
{"X_set": [["taster_twitter_handle", "@paulgwine "]], "Y_set": [["taster_name", "Paul Gregutt"]], "sup": 0.2667820990942994, "conf": 1.0, "lift": 3.7483774338492264}
{"X_set": [["taster_name", "Paul Gregutt"]], "Y_set": [["taster_twitter_handle", "@paulgwine "]], "sup": 0.2667820990942994, "conf": 1.0, "lift": 3.7483774338492264}
{"X_set": [["taster_twitter_handle", "@vboone"]], "Y_set": [["taster_name", "Virginie Boone"], ["province", "California"]], "sup": 0.2862280234416622, "conf": 1.0, "lift":
3.493718008375989}
{"X_set": [["taster_name", "Virginie Boone"]], "Y_set": [["province", "California"], ["taster_twitter_handle", "@vboone"]], "sup": 0.2862280234416622, "conf": 1.0, "lift":
3.493718008375989}
{"X_set": [["taster_twitter_handle", "@vboone"]], "Y_set": [["country", "US"], ["taster_name", "Virginie Boone"]], "sup": 0.2862280234416622, "conf": 1.0, "lift":
3.493718008375989}
{"X_set": [["taster_name", "Virginie Boone"]], "Y_set": [["country", "US"], ["taster_twitter_handle", "@vboone"]], "sup": 0.2862280234416622, "conf": 1.0, "lift":
3.493718008375989}
{"X_set": [["taster_twitter_handle", "@vboone"]], "Y_set": [["country", "US"], ["province", "California"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_name", "Virginie Boone"]], "Y_set": [["country", "US"], ["province", "California"]], "sup": 0.2862280234416622, "conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_name", "Paul Gregutt"]], "Y_set": [["country", "US"], ["taster_twitter_handle", "@paulgwine "]], "sup": 0.2667820990942994, "conf": 1.0, "lift":
3.7483774338492264}
{"X_set": [["taster_twitter_handle", "@paulgwine "]], "Y_set": [["country", "US"], ["taster_name", "Paul Gregutt"]], "sup": 0.2667820990942994, "conf": 1.0, "lift":
3.7483774338492264}
{"X_set": [["taster_name", "Virginie Boone"], ["taster_twitter_handle", "@vboone"]], "Y_set": [["country", "US"], ["province", "California"]], "sup": 0.2862280234416622,
"conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["country", "US"], ["taster_twitter_handle", "@vboone"]], "Y_set": [["taster_name", "Virginie Boone"], ["province", "California"]], "sup": 0.2862280234416622,
"conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["country", "US"], ["taster_name", "Virginie Boone"]], "Y_set": [["province", "California"], ["taster_twitter_handle", "@vboone"]], "sup": 0.2862280234416622,
"conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["province", "California"], ["taster_twitter_handle", "@vboone"]], "Y_set": [["country", "US"], ["taster_name", "Virginie Boone"]], "sup": 0.2862280234416622,
"conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_name", "Virginie Boone"], ["province", "California"]], "Y_set": [["country", "US"], ["taster_twitter_handle", "@vboone"]], "sup": 0.2862280234416622,
"conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_twitter_handle", "@vboone"]], "Y_set": [["country", "US"], ["taster_name", "Virginie Boone"], ["province", "California"]], "sup": 0.2862280234416622,
"conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["taster_name", "Virginie Boone"]], "Y_set": [["country", "US"], ["province", "California"], ["taster_twitter_handle", "@vboone"]], "sup": 0.2862280234416622,
"conf": 1.0, "lift": 3.493718008375989}
{"X_set": [["country", "US"]], "Y_set": [["province", "California"]], "sup": 0.5764961818504706, "conf": 0.5764961818504706, "lift": 0.5764961818504706}

本次实验通过 Apriori 算法按照 support, confidence, LIFT 等指标计算 出了频繁项集和他们之间的关系，对最终挖掘得到的关联规则进行分析我们 可以得知，"province"属性和"country"属性关联度极高，且 "province –> country"规则的置信度极高，与事实相符；还有"taster_twitter_handle"属性和 "taster_name"也有较大的关联度，及品酒者与其拥有的 twitter 账号名对应