Tìm tập mục thường xuyên và khai phá luật kết hợp

Phan Anh Phong, PhD.
Vinh University

Bài 3.1.

Cho chương trình sau:

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori
dataset = [['I1', 'I2', 'I5'],
           ['I2', 'I4'],
           ['I2', 'I3'],
           ['I1', 'I2', 'I4'],
           ['I1', 'I3'],
           ['I2', 'I3'],
           ['I1', 'I3'],
           ['I1', 'I2', 'I3', 'I5'],
           ['I1', 'I2', 'I3']]
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)
frequent_itemsets = apriori(df, min_support=2/9,
use_colnames=True)
print(frequent_itemsets)
```

1/ Cho biết chương trình trên thực hiện công việc gì. So sánh kết quả tìm được với cách tính thủ công

2/ Cho biết các giá trị trong biến te_ary

3/ Cho biết các tập mục thường xuyên với min_count = 3. So sánh kết quả tìm được với cách tính thủ công

4/ Bổ sung đoạn lệnh sau vào chương trình trên

```
from mlxtend.frequent_patterns import association_rules
rules=association_rules(frequent_itemsets,
metric="confidence", min_threshold=0.6)
print(rules.iloc[:, 0:3])
print(rules.iloc[:, 4:6])
```

a/ Cho biết kết quả thực hiện chương trình, giải thích chi tiết kết quả. So sánh kết quả tìm được với cách tính thủ công

b/ Thay đổi ngưỡng min_confidence thành 0.5. So sánh kết quả tìm được với cách tính thủ công

5/ Giả sử dữ liệu dưới đây được lưu trong 1 file .xls hoặc .csv với tên file data31.csv (data31.xls). Hãy chỉnh sửa chương trình trên để tìm tập mục thường xuyên và tập luật kết hợp. Viết báo cáo và nạp vào hệ thống LMS

```
dataset = [['Milk', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'],
['Dill', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'], ['Milk',
'Apple', 'Kidney Beans', 'Eggs'], ['Milk', 'Unicorn', 'Corn', 'Kidney
Beans', 'Yogurt'], ['Corn', 'Onion', 'Onion', 'Kidney Beans', 'Ice
cream', 'Eggs']]
```

Tham khảo thêm ở liên kết dưới đây:

http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/

6/ Giả sử cho data set ở liên kết này http://archive.ics.uci.edu/ml/machine-learning-databases/00352/

Hãy đề xuất giải pháp và thực hiện khai phá luật kết hợp trên tập dữ liệu đó. Gợi ý: Kết hợp nội dung ở Chương 2 và Chương 3

Yêu cầu: Viết báo cáo nạp vào hệ thống LMS

Bài 3.2.

1. Tạo file vidu1.cvs trên thư mục C:\datasets với nội dung như sau

Thit Bo	Thit Ga	Sua	
Thit Bo	Banh Fomat		
Banh Fomat	Giay		
Thit Bo	Thit Ga	Banh Fomat	
Thit Bo	Ao	Banh Fomat	Sua
Thit Ga	Quan Ao	Sua	
Thit Ga	Sua	Quan Ao	

2. Thực hiện chương trình

#Đoạn 1

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from apyori import apriori
# Đoạn 2
store_data = pd.read_csv('C:\\Datasets\\vidu1.csv',
header=None)
# Đoạn 3
records = []
for i in range(0, 7):
   records.append([str(store_data.values[i,j]) for j in
range(0, 2)])
# Đoạn 4
association_rules = apriori(records, min_support=0.15,
```

```
min_confidence=0.6, min_lift=3, min_length=2)
association_results = list(association_rules)
#Đoạn 5
for item in association_results:
    # first index of the inner list
    # Contains base item and add item
    pair = item[0]
    items = [x for x in pair]
    print("Rule: " + items[0] + " -> " + items[1])

    #second index of the inner list
    print("Support: " + str(item[1]))

    #third index of the list located at 0th
    #of the third index of the inner list

    print("Confidence: " + str(item[2][0][2]))
    print("Lift: " + str(item[2][0][3]))
    print("=====================================")
```

Giải thích: Lift(A→B) = (Confidence (A→B))/(Support (B))

2.1. Cho biết mỗi đoạn chương trình trên làm gì? Kiểm tra kết quả thu được với cách tính thủ công

2.2. Thay đổi giá trị của min_lift để hiểu hơn về ý nghĩa của thông số này

3. Khai phá luật kết hợp với chương trình trên cho dataset sau đây

dataset = [['Milk', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'],

['Dill', 'Onion', 'Nutmeg', 'Kidney Beans', 'Eggs', 'Yogurt'],

['Milk', 'Apple', 'Kidney Beans', 'Eggs'],

['Milk', 'Unicorn', 'Corn', 'Kidney Beans', 'Yogurt'],

['Corn', 'Onion', 'Onion', 'Kidney Beans', 'Ice cream', 'Eggs']]