BỘ GIÁO DỤC ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SÀI GÒN

Khoa Công nghệ Thông tin

PHÂN TÍCH DỮ LIỆU - BUỔI 06,7
Phân tích Luật kết hợp – Thực hành

CBGD: Phan Thành Huấn

✉ : pthuan112358@gmail.com

: 097 882 8842

Nội dung

1. Một số thư viện hỗ trợ phân tích luật kết hợp
2. Khai thác tập phổ biến
3. Khai thác luật kết hợp
4. Khai thác luật kết hợp dựa trên ràng buộc

1-Một số thư viện hỗ trợ phân tích luật kết hợp

Thư viện Python phổ biến được sử dụng để phân tích tập phổ biến và luật kết hợp:

1. mlxtend: Cung cấp một loạt các chức năng cho phân tích tập phổ biến và luật kết hợp; gồm thuật toán *Apriori* và *Eclat* cho phân tích tập phổ biến, cũng như thuật toán Association Rules cho LKH;
2. pyfpgrowth: Cung cấp thuật toán *FP-Growth* cho phân tích tập phổ biến và luật kết hợp - xây dựng trên cơ sở cấu trúc dữ liệu *FP-Tree*;

1-Một số thư viện hỗ trợ phân tích luật kết hợp

3. Orange: Cung cấp một loạt các công cụ và thuật toán cho phân tích tập phổ biến và luật kết hợp, bao gồm *Apriori* và *FP-Growth*;
4. pymining: Thư viện nhỏ gọn cho khai thác dữ liệu và LKH; cung cấp các hàm khai thác tập phổ biến và LKH. Hàm **frequent_itemsets()** để khai thác tập phổ biến và **rules()** cho khai thác LKH.

1-Một số thư viện hỗ trợ phân tích luật kết hợp

Một số hàm phổ biến trong thư viện mlxtend:

1. **apriori()**: Hàm này được sử dụng để áp dụng thuật toán *Apriori* để tìm tập phổ biến từ một *ma trận nhị phân*. Trả về một DataFrame chứa các tập phổ biến và độ phổ biến tương ứng;
2. **association_rules()**: Hàm này được sử dụng để áp dụng thuật toán *Association Rules* từ tập phổ biến đã tìm được. Trả về một DataFrame chứa các luật kết hợp và các thông số như độ tin cậy, độ phổ biến, lift, và leverage;

1-Một số thư viện hỗ trợ phân tích luật kết hợp

3. **TransactionEncoder()**: sử dụng để chuyển đổi dữ liệu thành *ma trận nhị phân*. Hàm sử dụng phương pháp *one-hot encoding* để chuyển đổi các mục thành các cột nhị phân;
4. **one_hot()**: chuyển đổi danh sách các thuộc tính thành một ma trận nhị phân. Trả về một *pandas DataFrame* có các cột nhị phân tương ứng với các thuộc tính trong danh sách;

1-Một số thư viện hỗ trợ phân tích luật kết hợp

5. **fpgrowth()**: sử dụng để áp dụng thuật toán FP-Growth để tìm tập phổ biến từ một *ma trận nhị phân*. Trả về một *DataFrame* chứa các tập phổ biến và độ phổ biến tương ứng;
6. **generate_association_rules()**: sử dụng để áp dụng thuật toán Association Rules từ tập phổ biến đã tìm được bằng FP-Growth. Nó trả về một *DataFrame* chứa các luật kết hợp và các thông số như độ tin cậy, độ phổ biến, lift, và leverage.

1-Một số thư viện hỗ trợ phân tích luật kết hợp

- Độ đo Lift là độ đo quan trọng trong phân tích luật kết hợp (association rules) và được sử dụng để đánh giá mức độ tương quan giữa các item trong một luật kết hợp.

$$lift(X \rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)\sup(Y)}$$

Giá trị Lift > 1 cho thấy mức độ tương quan tích cực giữa X và Y, trong khi giá trị Lift < 1 cho thấy mức độ tương quan tiêu cực. Nếu giá trị Lift = 1, thì không có mức độ tương quan giữa X và Y.

1-Một số thư viện hỗ trợ phân tích luật kết hợp

- Độ đo Leverage là độ đo trong phân tích LKH (association rules) được sử dụng để đánh giá mức độ tương quan giữa các item trong một LKH. Đo lường mức độ tương quan tuyến tính giữa việc xuất hiện của các item trong LKH so với việc chúng xuất hiện độc lập.

$$leverage(X \rightarrow Y) = \sup(X \cup Y) - \sup(X)\sup(Y)$$

Giá trị Leverage thuộc [-1, 1]. Giá trị Leverage = 0 không có tương quan tuyến tính giữa X và Y. Giá trị Leverage > 0 mức độ tương quan **dương**, trong khi Leverage < 0 mức độ tương quan **âm**.

1-Một số thư viện hỗ trợ phân tích luật kết hợp

Ví dụ: Sử dụng hàm apriori khai thác tập phổ biến thuộc mlxtend

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Dữ liệu mẫu
dataset = [['Bread', 'Milk', 'Eggs'],
           ['Bread', 'Diapers', 'Beer', 'Eggs'],
           ['Milk', 'Diapers', 'Beer', 'Coke'],
           ['Bread', 'Milk', 'Diapers', 'Beer'],
           ['Bread', 'Milk', 'Diapers', 'Coke']]

# Chuyển đổi dữ liệu thành ma trận nhị phân
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Áp dụng thuật toán Apriori để tìm tập phổ biến
frequent_itemsets = apriori(df, min_support=0.5, use_colnames=True)

# In ra tập phổ biến
print(frequent_itemsets)
```

	support	itemsets
0	0.6	(Beer)
1	0.8	(Bread)
2	0.8	(Diapers)
3	0.8	(Milk)
4	0.6	(Diapers, Beer)
5	0.6	(Diapers, Bread)
6	0.6	(Milk, Bread)
7	0.6	(Diapers, Milk)

1-Một số thư viện hỗ trợ phân tích luật kết hợp

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Dữ liệu mẫu
dataset = [['Bread', 'Milk', 'Eggs'],
           ['Bread', 'Diapers', 'Beer', 'Eggs'],
           ['Milk', 'Diapers', 'Beer', 'Coke'],
           ['Bread', 'Milk', 'Diapers', 'Beer'],
           ['Bread', 'Milk', 'Diapers', 'Coke']]

# Chuyển đổi dữ liệu thành ma trận nhị phân
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Áp dụng thuật toán Apriori để tìm tập phổ biến
frequent_itemsets = apriori(df, min_support=0.5, use_colnames=True)

# In ra tập phổ biến
print(frequent_itemsets)

# Áp dụng thuật toán Association Rules để tìm luật kết hợp
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.7)

# In ra các luật kết hợp
print(rules)
```

```
   support           itemsets
0      0.6             (Beer)
1      0.8            (Bread)
2      0.8          (Diapers)
3      0.8             (Milk)
4      0.6   (Diapers, Beer)
5      0.6  (Diapers, Bread)
6      0.6    (Milk, Bread)
7      0.6   (Diapers, Milk)
```

2-Khai thác tập phổ biến

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Tạo tập dữ liệu giả định
dataset = [['Milk', 'Bread', 'Butter'],
           ['Milk', 'Bread', 'Diapers'],
           ['Milk', 'Eggs', 'Butter'],
           ['Milk', 'Diapers', 'Butter'],
           ['Bread', 'Eggs', 'Butter']]

# Chuyển đổi dữ liệu thành ma trận nhị phân
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Áp dụng thuật toán Apriori để tìm tập phổ biến
frequent_itemsets = apriori(df, min_support=0.2, use_colnames=True)

# In kết quả
print(frequent_itemsets)
```

	support	itemsets
0	0.6	(Bread)
1	0.8	(Butter)
2	0.4	(Diapers)
3	0.4	(Eggs)
4	0.8	(Milk)
5	0.4	(Bread, Butter)
6	0.2	(Diapers, Bread)
7	0.2	(Eggs, Bread)
8	0.4	(Milk, Bread)
9	0.2	(Diapers, Butter)
10	0.4	(Eggs, Butter)
11	0.6	(Milk, Butter)
12	0.4	(Diapers, Milk)
13	0.2	(Eggs, Milk)
14	0.2	(Eggs, Bread, Butter)
15	0.2	(Milk, Bread, Butter)
16	0.2	(Diapers, Bread, Milk)
17	0.2	(Diapers, Butter, Milk)
18	0.2	(Eggs, Butter, Milk)

3-Khai thác luật kết hợp

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Tạo tập dữ liệu giả định
dataset = [['Milk', 'Bread', 'Butter'],
           ['Milk', 'Bread', 'Diapers'],
           ['Milk', 'Eggs', 'Butter'],
           ['Milk', 'Diapers', 'Butter'],
           ['Bread', 'Eggs', 'Butter']]

# Chuyển đổi dữ liệu thành ma trận nhị phân
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Áp dụng thuật toán Apriori để tìm tập phổ biến
frequent_itemsets = apriori(df, min_support=0.2, use_colnames=True)

# Áp dụng luật kết hợp
association_rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)

# In kết quả
print(association_rules)
```

3-Khai thác luật kết hợp

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

	antecedents	consequents	...	leverage	conviction
0	(Bread)	(Butter)	...	0.08	0.6
1	(Butter)	(Bread)	...	-0.08	0.8
2	(Diapers)	(Bread)	...	-0.04	0.8
3	(Eggs)	(Bread)	...	-0.04	0.8
4	(Bread)	(Milk)	...	-0.08	0.6
5	(Milk)	(Bread)	...	0.08	0.8
6	(Diapers)	(Butter)	...	-0.12	0.4
7	(Eggs)	(Butter)	...	0.08	inf
8	(Butter)	(Eggs)	...	0.08	1.2
9	(Butter)	(Milk)	...	-0.04	0.8
10	(Milk)	(Butter)	...	0.04	0.8
11	(Diapers)	(Milk)	...	0.08	inf
12	(Milk)	(Diapers)	...	0.08	1.2
13	(Eggs)	(Milk)	...	-0.12	0.4
14	(Bread, Butter)	(Eggs)	...	0.04	1.2
15	(Bread, Eggs)	(Butter)	...	0.04	inf
16	(Eggs, Butter)	(Bread)	...	-0.04	0.8
17	(Eggs)	(Bread, Butter)	...	0.04	1.2
18	(Bread, Butter)	(Milk)	...	-0.12	0.4
19	(Bread, Milk)	(Butter)	...	-0.12	0.4
20	(Bread, Milk)	(Diapers)	...	0.04	1.2
21	(Bread, Diapers)	(Milk)	...	0.04	inf
22	(Diapers, Milk)	(Bread)	...	-0.04	0.8
23	(Diapers)	(Bread, Milk)	...	0.04	1.2
24	(Diapers, Butter)	(Milk)	...	0.04	inf
25	(Diapers, Milk)	(Butter)	...	-0.12	0.4
26	(Diapers)	(Butter, Milk)	...	-0.04	0.8
27	(Eggs, Butter)	(Milk)	...	-0.12	0.4
28	(Eggs, Milk)	(Butter)	...	0.04	inf
29	(Eggs)	(Butter, Milk)	...	-0.04	0.8

3-Khai thác luật kết hợp

Ví dụ: Khai thác luật kết hợp từ “Chess.dat”

```
import numpy as np
import pandas as pd
import time
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Đường dẫn đến file dữ liệu
file_path = "chess.dat"

# Đọc dữ liệu từ file
df = np.loadtxt(file_path)

# Chuyển đổi dữ liệu thành ma trận nhị phân
te = TransactionEncoder()
te_ary = te.fit(df).transform(df)
df = pd.DataFrame(te_ary, columns=te.columns_)

start_time = time.time()
# Áp dụng thuật toán Apricri để tìm tập phổ biến
frequent_itemsets = apriori(df, min_support=0.8, use_colnames=True)
end_time = time.time()

# In ra tập phổ biến
print(frequent_itemsets)
print("Thoi gian khai thac tap pho bien:", end_time - start_time, "seconds")

start_time = time.time()
# Áp dụng thuật toán Association Rules để tìm luật kết hợp
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.8)
end_time = time.time()

# In ra các luật kết hợp
print(rules)
print("Thoi gian khai thac luat ket hop:", end_time - start_time, "seconds")
```

3-Khai thác luật kết hợp

Ví dụ: Khai thác luật kết hợp từ “Chess.dat”

```
        support                                                itemsets
0      0.888298                                                  (3.0)
1      0.929599                                                  (5.0)
2      0.962453                                                  (7.0)
3      0.899249                                                  (9.0)
4      0.894869                                                 (25.0)
...         ...                                                     ...
8222   0.801627  (66.0, 36.0, 40.0, 48.0, 52.0, 56.0, 58.0, 60....
8223   0.804130  (34.0, 66.0, 36.0, 7.0, 40.0, 48.0, 52.0, 58.0...
8224   0.801627  (36.0, 7.0, 40.0, 48.0, 52.0, 56.0, 58.0, 60.0...
8225   0.805069  (66.0, 36.0, 7.0, 40.0, 48.0, 52.0, 58.0, 60.0...
8226   0.803191  (34.0, 66.0, 36.0, 40.0, 48.0, 52.0, 58.0, 60....

[8227 rows x 2 columns]
Thoi gian khai thac tap pho bien: 2.3411340713500977 seconds
       antecedents  ... conviction
0              (3.0)  ...   0.999335
1              (5.0)  ...   0.999599
2              (3.0)  ...   1.440483
3              (7.0)  ...   1.104809
4              (9.0)  ...   1.059511
...              ...  ...        ...
552559      (52.0)  ...   0.999783
552560      (58.0)  ...   1.001279
552561      (60.0)  ...   1.064863
552562      (29.0)  ...   1.013138
552563      (62.0)  ...   1.007941

[552564 rows x 9 columns]
Thoi gian khai thac luat ket hop: 12.53171682357788 seconds
```

4-Khai thác luật kết hợp dựa trên ràng buộc

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Tạo tập dữ liệu giả định
dataset = [['Milk', 'Bread', 'Butter'],
           ['Milk', 'Bread', 'Diapers'],
           ['Milk', 'Eggs', 'Butter'],
           ['Milk', 'Diapers', 'Butter'],
           ['Bread', 'Eggs', 'Butter']]

# Chuyển đổi dữ liệu thành ma trận nhị phân
te = TransactionEncoder()
te_ary = te.fit(dataset).transform(dataset)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Áp dụng thuật toán Apriori để tìm tập phổ biến
frequent_itemsets = apriori(df, min_support=0.2, use_colnames=True)

# Áp dụng luật kết hợp
association_rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)

# Lọc các luật có chiều dài antecedents bằng 2
association_rules_2 = association_rules[association_rules['antecedents'].apply(lambda x: len(x) == 2)]

# In kết quả
print(association_rules_2)
```

4-Khai thác luật kết hợp dựa trên ràng buộc

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

	antecedents	consequents	...	leverage	conviction
0	(Bread)	(Butter)	...	-0.08	0.6
1	(Butter)	(Bread)	...	-0.08	0.8
2	(Diapers)	(Bread)	...	-0.04	0.8
3	(Eggs)	(Bread)	...	-0.04	0.8
4	(Bread)	(Milk)	...	0.08	0.6
5	(Milk)	(Bread)	...	-0.08	0.8
6	(Diapers)	(Butter)	...	-0.12	0.4
7	(Eggs)	(Butter)	...	0.08	1.6
8	(Butter)	(Eggs)	...	0.08	1.2
9	(Butter)	(Milk)	...	0.04	1.2
10	(Milk)	(Butter)	...	-0.04	0.8
11	(Diapers)	(Milk)	...	0.08	1.6
12	(Milk)	(Diapers)	...	0.08	1.2
13	(Eggs)	(Milk)	...	-0.12	0.4
14	(Bread, Butter)	(Eggs)	...	0.04	1.2
15	(Bread, Eggs)	(Butter)	...	0.04	inf
16	(Eggs, Butter)	(Bread)	...	-0.04	0.8
17	(Eggs)	(Bread, Butter)	...	0.04	1.2
18	(Bread, Butter)	(Milk)	...	-0.12	0.4
19	(Bread, Milk)	(Butter)	...	0.12	1.6
20	(Bread, Milk)	(Diapers)	...	0.04	1.2
21	(Bread, Diapers)	(Milk)	...	0.04	1.2
22	(Diapers, Milk)	(Bread)	...	-0.04	0.8
23	(Diapers)	(Bread, Milk)	...	0.04	1.2
24	(Diapers, Butter)	(Milk)	...	0.04	inf
25	(Diapers, Milk)	(Butter)	...	-0.12	0.4
26	(Diapers)	(Butter, Milk)	...	-0.04	0.8
27	(Eggs, Butter)	(Milk)	...	-0.12	0.4
28	(Eggs, Milk)	(Butter)	...	0.04	inf
29	(Eggs)	(Butter, Milk)	...	0.04	0.8

	antecedents	consequents	...	leverage	conviction
14	(Bread, Butter)	(Eggs)	...	0.04	1.2
15	(Bread, Eggs)	(Butter)	...	0.04	inf
16	(Butter, Eggs)	(Bread)	...	-0.04	0.8
18	(Milk, Bread)	(Butter)	...	-0.12	0.4
19	(Bread, Butter)	(Milk)	...	-0.12	0.4
20	(Milk, Diapers)	(Bread)	...	-0.04	0.8
21	(Milk, Bread)	(Diapers)	...	0.04	1.2
22	(Diapers, Bread)	(Milk)	...	0.04	inf
24	(Milk, Diapers)	(Butter)	...	-0.12	0.4
25	(Diapers, Butter)	(Milk)	...	0.04	inf
27	(Milk, Eggs)	(Butter)	...	0.04	inf
28	(Butter, Eggs)	(Milk)	...	-0.12	0.4

4-Khai thác luật kết hợp dựa trên ràng buộc

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

```
import numpy as np
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules

# Đường dẫn đến file dữ liệu
file_path = "chess.dat"

# Đọc dữ liệu từ file
df = np.loadtxt(file_path)

# Chuyển đổi dữ liệu thành ma trận nhị phân
te = TransactionEncoder()
te_ary = te.fit(df).transform(df)
df = pd.DataFrame(te_ary, columns=te.columns_)

# Áp dụng thuật toán Apriori để tìm tập phổ biến
frequent_itemsets = apriori(df, min_support=0.8, use_colnames=True)

# In ra tập phổ biến
print(frequent_itemsets)

# Áp dụng thuật toán Association Rules để tìm luật kết hợp
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.8)

# In ra các luật kết hợp
print(rules)
```

4-Khai thác luật kết hợp dựa trên ràng buộc

Ví dụ: Sử dụng hàm apriori và association_rules thuộc mlxtend

```
        support                                               itemsets
0      0.888298                                                (3.0)
1      0.929599                                                (5.0)
2      0.962453                                                (7.0)
3      0.899249                                                (9.0)
4      0.894869                                               (25.0)
...         ...                                                   ...
8222   0.801627  (66.0, 36.0, 40.0, 48.0, 52.0, 56.0, 58.0, 60....
8223   0.804130  (34.0, 66.0, 36.0, 7.0, 40.0, 48.0, 52.0, 58.0...
8224   0.801627  (36.0, 7.0, 40.0, 48.0, 52.0, 56.0, 58.0, 60.0...
8225   0.805069  (66.0, 36.0, 7.0, 40.0, 48.0, 52.0, 58.0, 60.0...
8226   0.803191  (34.0, 66.0, 36.0, 40.0, 48.0, 52.0, 58.0, 60....

[8227 rows x 2 columns]
       antecedents  ... conviction
0            (3.0)  ...   0.999335
1            (5.0)  ...   0.999599
2            (3.0)  ...   1.440483
3            (7.0)  ...   1.104809
4            (9.0)  ...   1.059511
...            ...  ...        ...
552559      (52.0)  ...   0.999783
552560      (58.0)  ...   1.001279
552561      (60.0)  ...   1.064863
552562      (29.0)  ...   1.013138
552563      (62.0)  ...   1.007941

[552564 rows x 9 columns]
```