

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Association Rules Report



Bộ môn Khoa học Máy tính
Khoa Công nghệ thông tin
Đại học Khoa học tự nhiên TP HCM

Mục lục

1	Thông tin nhóm	3
2	Apriori.....	4
3	FP-Growth	6
4	Tập phổ biến đóng và tập phổ biến tối đại	9
5	Các độ đo lý thú.....	10

1

Thông tin nhóm

MSSV	Họ Tên	Email
1512203	Nguyễn Quốc Huy	1512203@student.hcmus.edu.vn
1512262	Võ Anh Khoa	1512262@student.hcmus.edu.vn

2 Apriori

Các bước tìm các frequent itemsets với minsup = 60% (min.count ≥ 4):

Bước 1:

C1	Support
Bia	3
Bánh mì	4
Hành	4
Sữa	4
Khoai tây	5

L1	Support
Sữa	4
Hành	4
Bánh mì	4
Khoai tây	5


Bước 2

C2	Support
Sữa, Hành	2
Sữa, Bánh mì	2
Sữa, Khoai tây	3
Hành, Bánh mì	3
Khoai tây, Hành	4
Khoai tây, Bánh mì	4

L2	Support
Khoai tây, Hành	4
Khoai tây, Bánh mì	4

Bước 3

C3	Support
Khoai tây, Hành, Bánh mì	3

 : Các itemsets có support < minsup

Các frequent itemsets:

{Sữa}, {Khoai tây}, {Hành}, {Bánh mì}, {Khoai tây, Hành}, {Khoai tây, Bánh mì}

Bước 4: Tính confident, min conf = 80%

Association Rules	Support(A,B)	Support(A)	Confidence
Bánh mì \Rightarrow Khoai tây	4	4	100%
Hành \Rightarrow Khoai tây	4	4	100%
Khoai tây \Rightarrow Bánh mì	4	5	80%
Khoai tây \Rightarrow Hành	4	5	80%

Vậy ta tìm được 4 luật kết hợp thỏa mãn $\text{minsup} \geq 60\%$ và $\text{minconf} \geq 80\%$:

R1: Khoai tây \Rightarrow Hành (support = 67%, confidence = 80%)

R1: Hành \Rightarrow Khoai tây (support = 67%, confidence = 100%)

R1: Khoai tây \Rightarrow Bánh mì (support = 67%, confidence = 80%)

R1: Bánh mì \Rightarrow Khoai tây (support = 67%, confidence = 100%)

3

FP-Growth

3.1 Mô tả source code

Class *FPTreeOperationContainer* dùng để quản lý FP-Tree:

```
class FPTreeOperationContainer {
private:
    // tạo nút mới
    static FPTreeNode newFPTreeNode(int itemID, FPTreeNode parent = NULL);

    // từ nút p trên cây, tìm nút con của p mà có ID của item là itemID
    FPTreeNode findBranchToGo(FPTreeNode p, int itemID);
    // từ nút p trên cây, tìm nút con của p mà có ID của item là itemID, nếu không tồn tại thì
    // tạo ra nút mới
    FPTreeNode makeNewConnection(int itemID, FPTreeNode p);
    // sinh ra conditional FPTree cho item có ID là itemID
    FPTreeOperationContainer* unblockConditionalFPTree(int itemID);

    // gốc cây
    FPTreeNode root;
    // Số item, số transaction
    int nItems, nTransactions;
    // danh sách các item được sắp xếp với tần số giảm dần
    int *itemOrdered;
    // danh sách các nút của cây (quản lý theo từng item)
    std::vector<FPTreeNode> *headList;
public:
    FPTreeOperationContainer(int _nItems, int* _itemOrdered);
    ~FPTreeOperationContainer();

    // thêm một transaction (biết tần số)
    void insertTransaction(const std::vector<int> &transaction, int freq = 1);
    // sinh tập phổ biến từ cây (biết ngưỡng min support)
    std::vector< std::vector<int> > findConditionalFrequentSet(double threshold);
```

};

Mã giả sinh tập phổ biến từ cây FP:

Sinh-Tập-Phổ-Biến(*Cây-fp*, *minSup*) {
 Kết-quả = {} // Rỗng

Xét các item *iid* theo thứ tự tần số tăng dần {
Cây-Cfp = Sinh-conditional-FP-Tree(*Cây-fp*, *iid*)
 Nếu Số-transaction(*Cây-Cfp*) < *minSup* {
 continue;
 }
Tập-con = Sinh-Tập-Phổ-Biến(*Cây-Cfp*, *minSup*);
Tập-con = *Tập-con* \cup { {} }
 Với mọi phần tử *itemset* thuộc tập con {
 itemset = *itemset* \cup { *iid* }
 Kết-quả = *Kết-quả* \cup { *itemset* }
 }
 }
 return *Kết-quả*
}

Sinh-conditional-FP-Tree(*Cây-fp*, *iid*) {
 Count[*i*] = 0 với mọi *i* thuộc {0, ..., Số-item(*Cây-fp*)-1}
 Với mọi nút *p* của *Cây-fp* có itemID bằng *iid* {
 itemlist = Danh sách item khi duyệt đường đi từ *p* lên Nút-gốc(*Cây-fp*)
 Với mọi *i* thuộc *itemlist* {
 Count[Item-ID(*i*)] += Frequency(*i*)
 }
 }
Cây-kết-quả = Cây-Rỗng()
 Với mọi nút *p* của *Cây-fp* có itemID bằng *iid* {
 itemlist = Danh sách item khi duyệt đường đi từ *p* lên Nút-gốc(*Cây-fp*)
 Sắp xếp lại *itemlist* theo Count giảm dần
 Chèn-transaction(*Cây-kết-quả*, *itemlist*, Frequency(*i*))
 }
 return *Cây-kết-quả*

}

3.2 So sánh thuật toán Apriori và FP-Growth

Apriori	FPTree
Phải sinh tập ứng viên (candidate set) lớn	Không phải sinh tập ứng viên (candidate set) lớn
Phải duyệt lại database nhiều lần	Chỉ phải duyệt database ban đầu 1 lần
Chỉ phải tốn bộ nhớ để lưu CSDL và tập ứng viên	Phải phát sinh conditional tree một cách đệ quy -> tốn bộ nhớ

4 Tập phổ biến đóng và tập phổ biến tối đại

4.1 Tập phổ biến đóng

Tập phổ biến đóng là tập phổ biến mà không có tập nào bao nó có cùng độ phổ biến.

$$F = \{X \mid X \subseteq I \wedge \text{sup}(X) \geq \text{minsup}\}$$

- Với F là tập hợp gồm tất cả tập phổ biến.
- Gọi C là tập hợp gồm tất cả tập phổ biến đóng.

$$C = \{X \mid X \in F \wedge \nexists Y \supset X, \text{sup}(X) = \text{sup}(Y)\}$$

4.2 Tập phổ biến tối đại

Tập phổ biến tối đại là tập phổ biến mà không có tập nào bao nó là phổ biến.

$$C = \{X \mid X \in F \wedge \nexists (Y \supset X \mid Y \in F)\}$$

4.3 Tập phổ biến đóng và tối đại cho bài 1

Tập phổ biến đóng:

{Sữa}, {Khoai tây}, {Khoai tây, Hành}, {Khoai tây, Bánh mì}.

Tập phổ biến tối đại:

{Sữa}, {Khoai tây, Hành}, {Khoai tây, Bánh mì}.

5 Các độ đo lý thú

5.1 Công thức tính các độ đo confidence, lift, conviction, leverage

Độ đo Confidence trình bày mức độ thường xuyên của tính đúng đắn của luật

Độ đo của luật $X \Rightarrow Y$, ứng theo tập hợp giao tác T là tỉ lệ giao tác có chứa X và Y so với giao tác chỉ chứa X.

Độ đo Confidence được định nghĩa như sau:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

Độ đo Lift là tỷ lệ của sự hỗ trợ quan sát được đối với X và Y là độc lập

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

Độ đo Conviction được định nghĩa:

$$\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$$

Độ đo Leverage được định nghĩa

$$\text{leverage}(X \Rightarrow Y) = \text{supp}(X \Rightarrow Y) - \text{supp}(X)\text{supp}(Y)$$

5.2 Tính các độ đo cho các luật kết hợp ở bài 1

	Lift	Conviction	Leverage
Khoai tây \Rightarrow Hành	1.2	1.67	0.11
Hành \Rightarrow Khoai tây	1.2	∞	0.11
Khoai tây \Rightarrow Bánh mì	1.2	1.67	0.11
Bánh mì \Rightarrow Khoai tây	1.2	∞	0.11

- Đối với Luật : *Khoai Tây* \Rightarrow *Hành*

Trước hết ta tính $\text{supp}(\text{Khoai Tây})$, $\text{supp}(\text{Hành})$, $\text{supp}(\text{Khoai Tây} \cup \text{Hành})$

Ta có :

$$\text{supp}(\text{Khoai Tây}) = \frac{5}{6}, \text{supp}(\text{Hành}) = \frac{2}{3}, \text{supp}(\text{Khoai Tây} \cup \text{Hành}) = \frac{2}{3}$$

$$\text{conf}(\text{Khoai Tây} \Rightarrow \text{Hành}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Hành})}{\text{supp}(\text{Khoai Tây})} = \frac{\frac{2}{3}}{\frac{5}{6}} = \frac{4}{5}$$

$$\text{lift}(\text{Khoai Tây} \Rightarrow \text{Hành}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Hành})}{\text{supp}(\text{Khoai Tây}) \times \text{supp}(\text{Hành})} = \frac{\frac{2}{3}}{\frac{5}{6} \times \frac{2}{3}} = \frac{6}{5}$$

$$\text{conv}(\text{Khoai Tây} \Rightarrow \text{Hành}) = \frac{1 - \text{supp}(\text{Hành})}{1 - \text{conf}(\text{Khoai Tây} \Rightarrow \text{Hành})} = \frac{1 - \frac{2}{3}}{1 - \frac{4}{5}} = \frac{5}{3}$$

$\text{leverage}(\text{Khoai Tây} \Rightarrow \text{Hành})$

$$= \text{supp}(\text{Khoai Tây} \Rightarrow \text{Hành}) - \text{supp}(\text{Khoai Tây})\text{supp}(\text{Hành})$$

$$= \frac{2}{3} - \frac{2}{3} \times \frac{5}{6} = \frac{1}{9}$$

- Đối với Luật : *Hành* \Rightarrow *Khoai Tây*

Trước hết ta tính $\text{supp}(\text{Khoai Tây})$, $\text{supp}(\text{Hành})$, $\text{supp}(\text{Khoai Tây} \cup \text{Hành})$

Ta có :

$$\text{supp}(\text{Khoai Tây}) = \frac{5}{6}, \text{supp}(\text{Hành}) = \frac{2}{3}, \text{supp}(\text{Khoai Tây} \cup \text{Hành}) = \frac{2}{3}$$

$$\text{conf}(\text{Hành} \Rightarrow \text{Khoai Tây}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Hành})}{\text{supp}(\text{Hành})} = \frac{\frac{2}{3}}{\frac{2}{3}} = 1$$

$$\text{lift}(\text{Hành} \Rightarrow \text{Khoai Tây}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Hành})}{\text{supp}(\text{Khoai Tây}) \times \text{supp}(\text{Hành})} = \frac{\frac{2}{3}}{\frac{5}{6} \times \frac{2}{3}} = \frac{6}{5}$$

$$\text{conv}(\text{Hành} \Rightarrow \text{Khoai Tây}) = \frac{1 - \text{supp}(\text{Khoai Tây})}{1 - \text{conf}(\text{Hành} \Rightarrow \text{Khoai Tây})} = \frac{1 - \frac{5}{6}}{1 - 1} = \infty$$

leverage(Hành \Rightarrow Khoai Tây)

$$= \text{supp}(\text{Hành} \Rightarrow \text{Khoai Tây}) - \text{supp}(\text{Khoai Tây})\text{supp}(\text{Hành}) = \frac{2}{3} - \frac{2}{3} \times \frac{5}{6} \\ = \frac{1}{9}$$

- Đối với Luật : *Khoai Tây \Rightarrow Bánh mì*

Trước hết ta tính $\text{supp}(\text{Khoai Tây})$, $\text{supp}(\text{Bánh Mì})$, $\text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì})$

Ta có :

$$\text{supp}(\text{Khoai Tây}) = \frac{5}{6}, \text{supp}(\text{Bánh Mì}) = \frac{2}{3}, \text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì}) = \frac{2}{3}$$

$$\text{conf}(\text{Khoai Tây} \Rightarrow \text{Bánh mì}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì})}{\text{supp}(\text{Khoai Tây})} = \frac{\frac{2}{3}}{\frac{5}{6}} = \frac{4}{5}$$

$$\text{lift}(\text{Khoai Tây} \Rightarrow \text{Bánh mì}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì})}{\text{supp}(\text{Khoai Tây}) \times \text{supp}(\text{Bánh Mì})} = \frac{\frac{2}{3}}{\frac{5}{6} \times \frac{2}{3}} = \frac{6}{5}$$

$$\text{conv}(\text{Khoai Tây} \Rightarrow \text{Bánh Mì}) = \frac{1 - \text{supp}(\text{Bánh Mì})}{1 - \text{conf}(\text{Khoai Tây} \Rightarrow \text{Bánh Mì})} = \frac{1 - \frac{2}{3}}{1 - \frac{4}{5}} = \frac{5}{3}$$

leverage(*Khoai Tây \Rightarrow Bánh Mì*)

$$= \text{supp}(\text{Khoai Tây} \Rightarrow \text{Bánh Mì}) - \text{supp}(\text{Khoai Tây})\text{supp}(\text{Bánh Mì}) \\ = \frac{2}{3} - \frac{2}{3} \times \frac{5}{6} = \frac{1}{9}$$

- Đối với Luật : *Bánh Mì \Rightarrow Khoai Tây*

Trước hết ta tính $\text{supp}(\text{Khoai Tây})$, $\text{supp}(\text{Bánh Mì})$, $\text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì})$

Ta có :

$$\text{supp}(\text{Khoai Tây}) = \frac{5}{6}, \text{supp}(\text{Bánh Mì}) = \frac{2}{3}, \text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì}) = \frac{2}{3}$$

$$\text{conf}(\text{Bánh Mì} \Rightarrow \text{Khoai Tây}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì})}{\text{supp}(\text{Bánh Mì})} = \frac{\frac{2}{3}}{\frac{2}{3}} = 1$$

$$\text{lift}(\text{Bánh Mì} \Rightarrow \text{Khoai Tây}) = \frac{\text{supp}(\text{Khoai Tây} \cup \text{Bánh Mì})}{\text{supp}(\text{Khoai Tây}) \times \text{supp}(\text{Bánh Mì})} = \frac{\frac{2}{3}}{\frac{5}{6} \times \frac{2}{3}} = \frac{6}{5}$$

$$\text{conv}(\text{Bánh Mì} \Rightarrow \text{Khoai Tây}) = \frac{1 - \text{supp}(\text{Khoai Tây})}{1 - \text{conf}(\text{Bánh Mì} \Rightarrow \text{Khoai Tây})} = \frac{1 - \frac{5}{6}}{1 - 1} = \infty$$

$$\begin{aligned} \text{leverage}(\text{Bánh Mì} \Rightarrow \text{Khoai Tây}) &= \text{supp}(\text{Bánh Mì} \Rightarrow \text{Khoai Tây}) - \text{supp}(\text{Khoai Tây})\text{supp}(\text{Bánh Mì}) \\ &= \frac{2}{3} - \frac{2}{3} \times \frac{5}{6} = \frac{1}{9} \end{aligned}$$

5.3 Nhận xét

Các độ đo này có sự khác biệt, vì ý nghĩa của chúng có sự khác biệt.

Confidence: Confidence cho biết tỉ lệ phần trăm một luật tìm được là đúng.

Lift: Lift cho biết độ độc lập giữa 2 transaction. Nếu lift xấp xỉ với 1 thì A và B độc lập. Nếu $\text{lift} > 1$, giá trị của nó cho ta biết mức độ phụ thuộc giữa 2 transaction, và dẫn đến tăng tính hữu ích. Thường hiệu quả với những item có support nhỏ.

Conviction: Tương tự như Lift, nhưng khắc phục được nhược điểm của lift là không phân biệt được chiều của luật.

Leverage: Leverage cho biết mức độ độc lập giữa 2 items. Thường hiệu quả với những item có support lớn.