A Data Mining Framework for Optimal Product Selection

K. Vanhoof and T.Brijs

Department of Applied Economic Sciences

Limburg University Centre

Belgium

Outline

- Research setting
- Market basket analysis
- Model for product selection : Profset
- Problems and other approaches
- Generalized Profset
- Conclusions

Research Setting

Use product interdependency information to
construct an optimisation **model** to select a
hitlist of products
that yields **maximum profits**
based on their cross-selling effects
incorporating business constraints
and compare with product profitability heuristic

Market Basket Analysis

Let

- $I = \{i_1, i_2, \ldots, i_k\}$ be the product assortment of the retail store
- T be a basket of products such that $T \subseteq I$
- D be a database of baskets T
- X be a set of items such that $X \subseteq I$

– Definition 1: support of an itemset

  $S(X, D)$ = the support of an itemset X in D, i.e. the fraction of transactions in D that contain X

– Definition 2: Frequent itemset

- An itemset is called frequent in D, if its support exceeds a user-predefined threshold called *minimum support*

Model for Product Selection

- Objective: Suppose you have 35 eye-level facings in the convenience store: determine which set of products yields maximum profits, taking into account cross-selling effects

Model for Product Selection

- Criteria that influence profitability
  - Product margin
  - Product handling cost (restocking)
  - Product inventory costs (refrigerating, financial)
- Degrees of freedom
  - Number of facings (size of the hitlist)
- Model constraints
  - Which categories <u>must</u> be included
  - How many products of each category
  - Specify base products (e.g. to support store image)

Model for Product Selection

- $m(T)$: gross margin generated by transaction T

$$m(T) = \sum_{i \in T} (SP(i) - PP(i)) * f(i)$$

- $M(X)$: gross margin generated by frequent itemset X

$$M(X) = \sum_{T \in D} m'(T) \quad \text{with} \quad \begin{cases} m'(T) = m(T) \text{ if } X = T \\ m'(T) = 0 \text{ otherwise} \end{cases}$$

Model for Product Selection

- PROFSET model
  - Let
    - L be the set of all frequent itemsets
    - Categories $C_1, \ldots, C_n$ be sets of items
    - $P_X, Q_i \in \{0,1\}$ decision variables to be optimized
    - $Cost_i$ be total cost (storage + handling) per product i

Model for Product Selection

- Objective function

$$Max\left(\sum_{X \in L} M(X)\, P_X - \sum_{c=1}^{n} \sum_{i \in C_c} Cost_i\, Q_i\right)$$

- Constraints

$$\sum_{c=1}^{n} \sum_{i \in C_c} Q_i = ItemMax \qquad (1)$$

$$\forall X \in L, \forall i \in X : Q_i \geq P_X \qquad (2)$$

$$\forall C_c : \sum_{i \in C_c} Q_i \geq ItemMin_{C_c} \qquad (3)$$

Model for Product Selection

- Suppose
  - X_1 = {cola, peanuts} $M(X_1)$=10
  - X_2 = {peanuts, cheese} $M(X_2)$=20
  - X_3 = {peanuts, beer} $M(X_3)$=30
  - $Cost_{cola}$=5, $Cost_{peanuts}$=3, $Cost_{cheese}$=1, $Cost_{beer}$=4
  - You are allowed to select only 3 items, which ones would you choose?

Model for Product Selection

$$Z = 10P_1 + 20P_2 + 30P_3 - 5Q_1 - 3Q_2 - 1Q_3 - 4Q_4$$

s.t.

$$Q_1 + Q_2 + Q_3 + Q_4 \leq 3 \qquad (1)$$

$$Q_1 \geq P_1$$

$$Q_2 \geq P_1$$

$$Q_2 \geq P_2 \qquad (2)$$

$$Q_3 \geq P_2$$

$$Q_2 \geq P_3$$

$$Q_4 \geq P_3$$

Model for Product Selection

$$Z = 10P_1 + 20P_2 + 30(1) - 5Q_1 - 3(1) - 1Q_3 - 4(1)$$

s.t.

$$Q_1 + 1 + Q_3 + 1 \leq 3 \qquad (1)$$

$$Q_1 \geq P_1$$

$$1 \geq P_1$$

$$1 \geq P_2 \qquad (2)$$

$$Q_3 \geq P_2$$

$$1 \geq 1$$

$$1 \geq 1$$

Model for Product Selection

$$Z = 10P_1 + 20(1) + 30(1) - 5Q_1 - 3(1) - 1(1) - 4(1)$$

s.t.

$$Q_1 + 1 + 1 + 1 \leq 3 \qquad (1)$$

$$Q_1 \geq P_1$$

$$1 \geq P_1$$

$$1 \geq 1 \qquad (2)$$

$$1 \geq 1$$

$$1 \geq 1$$

$$1 \geq 1$$

cola = 0

Peanuts, cheese, beer = 1

Z = 42

Remarks

- $m'(T) = 0$ for $X \subset T$

  baskets containing infrequent products are out the model

– When one product x of a frequent set is not selected the total margin of the frequent set reduces to zero , even when a substitute of x is available

– Key requirement : average size of basket $\cong$

  size of largest frequent itemsets

Models with loss rule

- Margin of product x in selection S is corrected for not selected products

  $T = T_S \cup D$
- $m_S(x) = m(x)\ (\ 1 - csf\ (D,x)\ )$
- csf is the cross selling factor
- Total Profit = Max $\sum\sum m_S(x)$

One approach (Wong,Fu,Wang)

- If we have a rule that purchasing product x implies buying a product of D then the margin of T should be zero.
- The higher the confidence of that rule the more likely the margin should decrease
- Csf = conf ($x \Rightarrow \exists D$)

Remarks

- Suppose product x and set D are independent

Cross selling factor should be zero or low

but : conf (x $\Rightarrow \exists$ D) $\approx$ supp ($\exists$ D)

and supp ($\exists$ D) can be large for large D

Another approach (Wang,Su)

- If we have a rule that product x will not be purchased due to the absence of products D then the margin of x a should be zero.
- The higher the confidence of that rule the more likely the margin should decrease
- Csf = conf (not D $\Rightarrow$ not x)

Remarks

- Suppose product x and set D are independent

  Cross selling factor should be zero

  but : conf (not D $\Rightarrow$ not x) $\approx$ supp (not x)

  with supp (not x) very high

Gross Margin per Frequent Itemset for Small Market Baskets

- Key observation: average size of basket ≅ size of largest frequent itemset

- Gross margin per transaction T
  - m(T) = $\sum_{i \in T} (SP(i) - PP(i)) * f(i)$

- Gross margin per frequent itemset X
  - M(X) = $\sum_{T \in D} m'(T)$ with $\begin{cases} m'(T) = m(T) & \text{if } X = T \\ m'(T) = 0 & \text{if } X \neq T \end{cases}$

Generalised Profset approach

- Key observation: average size of market basket >> size of largest frequent itemset
- How to distribute $m'(T)$ over multiple frequent subsets of T?
- Idea: Allocate the proportion of $m'(T)$ to the purchase combination that was most likely the purchase intention of the consumer at the time of purchase
  - We can't know exactly: one should have asked the consumer
  - But, we can estimate it in terms of probabilities

Gross Margin per Frequent Itemset for Large Market Baskets

- Definition: Maximal frequent subset of a transaction
  - Let F be the collection of all frequent subsets of T.
    Then $X \in F$ is called *maximal*, denoted as X_{max}, if and only if $\forall Y \in F: |Y| \leq |X|$
  - E.g. T = {cola, peanuts, cheese}

Frequent Sets	**Support**	**Maximal**	**Unique**
{cola}	10%	No	No
{peanuts}	5%	No	No
{cheese}	8%	No	No
{cola, peanuts}	2%	Yes	No
{peanuts, cheese}	1%	Yes	No

Gross Margin per Frequent Itemset for Large Market Baskets

- Profit Allocation Algorithm

For every transaction T **do** {
    while (T contains frequent sets) **do** {
        Draw X from all maximal frequent subsets using probability distribution Θ_T;
        $M(X) := M(X) + m(X)$
        with $m(X)$ the profit margin of X in T;
        $T := T \setminus X$;
    }
}
return all $M(X)$;

- Probability distribution Θ_T

$$\Theta_T(X_{max}) = \sup(X_{max}) \; / \sum_{Y_{\max} \in T} \sup(Y_{\max})$$

Example

E.g. T = {cola, peanuts, cheese}
with $m(T)$ = 1\$ + 1.5\$ + 2\$ = 4.5\$

Frequent Sets	Support	Maximal	Unique
{cola}	10%	No	No
{peanuts}	5%	No	No
{cheese}	8%	No	No
{cola, peanuts}	2%	Yes	No
{peanuts, cheese}	1%	Yes	No

We draw {cola, peanuts} with probability 2/3 and {peanuts, cheese} with probability 1/3 from probability distribution $\Theta_T(X_{max})$.

Suppose, {cola, peanuts} is drawn, then M(cola, peanuts):= M(cola,peanuts) = 2.5\$ and $T := T \setminus X$, i.e. T = {cheese}

Finally, M(cheese):=M(cheese)=2\$, and $T = \{\emptyset\}$

The Generalized PROFSET Model

- Let
  - C_1 , C_2 , …, C_n be sets of items (categories)
  - L be the set of frequent sets
  - P_X, $Q_i \in \{0,1\}$ the decision variables
  - $Cost_i$ be the inventory and handling cost of product i
- Max $\left(\sum_{X \in L} M(X) P_X - \sum_{c=1}^{n} \sum_{i \in C_c} Cost_i Q_i \right)$

$$\sum_{c=1}^{n} \sum_{i \in C_c} Q_i = ItemMax \qquad (1)$$

$$\forall X \in L, \forall i \in X : Q_i \geq P_X \qquad (2)$$

$$\forall C_c : \sum_{i \in C_c} Q_i \geq ItemMin_{C_c} \qquad (3)$$

Empirical Study

- Belgian supermarket store
  - 18182 market baskets with average size of 10.6
  - 9965 different products
  - 281 product categories
  - 3381 consumers with loyalty card
  - 87% of assortment is slow moving, i.e. sold less than once a day

- Objective
  - Determine the set of eye-catcher products that yields maximum profits from cross-selling. However, because of limited shelf-space, only one delegate product per category may be selected.

Empirical Study

- Results
  - For 25% of the categories, PROFSET selects a different product compared to the one with the highest profit ranking within each category. This must be due to cross-selling effects between eye-catcher products.
  - Profit improvements per category were between 3% and 588%

Product	Own profit (BEF)	Cross-selling profit (BEF)	Total profit (BEF)
Milky Way Mini	37808	2350	40158
Melo Cakes	34333	0	34333
Leo 3-pack	28728	0	28728
Leo 10 + 2 pack	12028	264228	276256

  - Why?

Empirical Study

- People tend to buy other eye-catcher products with Leo 10+2 pack. Looking at the package size (10+2) this is not surprising since it is targeted at larger families.
- Validation by means of association rules: consumers who buy Leo 10+2 pack tend to do one-stop shopping whereas Milky Way is a single item snack and does not sell so often with other products:
  - Milky Way Mini $\Rightarrow$ Vegetable/Fruit (sup=0.17%, conf=50.82%)
  - Meat product and Leo 10+2 pack $\Rightarrow$ cheese product (sup=0.396%, conf=55%)

Conclusions and further research

- **Conclusion**
  - The generalized PROFSET model for large market baskets adequatly solves the problem of product selection in big supermarkets.
  - PROFSET can be easily adapted to reflect a number of category management constraints imposed by the retailer and thus provides wide flexibility to incorporate retail domain knowledge.

- **Further research**
  - Alternative profit allocation schemes and their impact on the selection of an optimal set by PROFSET.
  - Validation in real world situations.

References

- MPIS : Maximal-Profit Item selection with cross-selling considerations
  Chi-Wing Wong, Ada Wai-Chee Fu, Ke Wang
- Building an Association rules Framework to Improve Product Assortment decisions Data mining and knowledge discovery, 8, 7-23,2004
  Tom Brijs, Gilbert Swinnen, Koen Vanhoof, Geert Wets
- A data mining framework for optimal product selection in retail supermarket data: the generalised profset model SIGKDD 2000
  Tom Brijs, Bart Goethals, Gilbert Swinnen, Koen Vanhoof, Geert Wets
- Using associations rules for product assortment decisions. A case study. SIGKDD 1999
  Tom Brijs, Gilbert Swinnen, Koen Vanhoof, Geert Wets
- Item selection by ‘hub-authority’ profit ranking. SIGKDD 2002
  Wang K. , Su M.Y.