MINING ASSOCIATION RULES

STQD6414 PERLOMBONGAN DATA

Assoc. Prof. Dr. Nurulkamal Masseran

Department of Mathematical Sciences

Universiti Kebangsaan Malaysia

INTRODUCTION:

- The Mining Association Rule aims to find ‘interesting’ relationships between sets of items.
- This technique is commonly used to make product recommendations by identifying products that are often purchased together.
- Mining Association Rule is typically performed on transaction data from retail markets or from online e-commerce businesses.
- A priori and eclat algorithms are used to find a patterns and rules in the dataset.

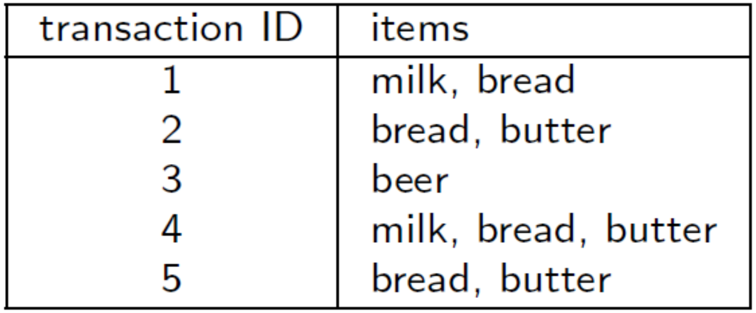
- What is rule?
- Rules refer to a notations that represent items that are often purchased along with other items.
- It has the LHS and RHS parts represented as follows:

$$X \Rightarrow Y$$

- That means, the item on the right is often purchased along with the item on the left.

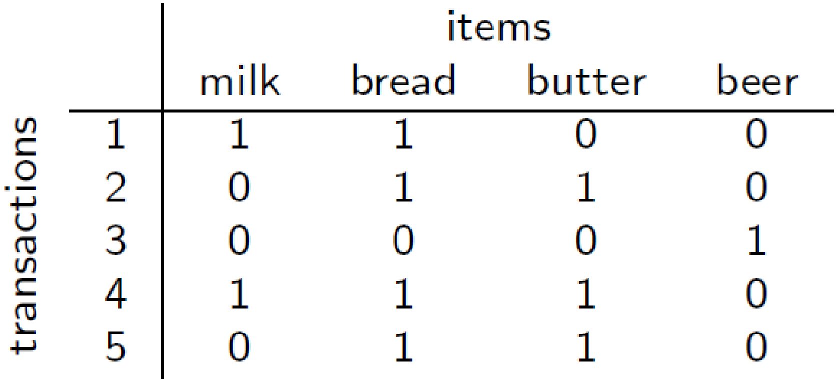
DATA TRANSACTIONS:

- Example of Product Purchasing Transaction Data :



transaction ID	items
1	milk, bread
2	bread, butter
3	beer
4	milk, bread, butter
5	bread, butter

- Data mining algorithm reads the transaction data in the form of binary variables
- Where $I=\{i_1, i_2, \ldots, i_n\}$ is a set of *n*-binary attributes referred to as items and $D=\{t_1, t_2, \ldots, t_m\}$ is the set of *m*-transactions.



		items			
		milk	bread	butter	beer
transactions	1	1	1	0	0
	2	0	1	1	0
	3	0	0	0	1
	4	1	1	1	0
	5	0	1	1	0

BASIC OF ASSOCIATION RULES:

- Association rules represent by $X \Rightarrow Y$.
- On condition:

i) $X, Y \subseteq I$

ii) $X \cap Y = \varnothing$ (X and Y are not the same items).

iii) X is the antecedent rule (events that occur first).

iv) Y is a consequential rule (an event that occurs due to something).

Example: **{Milk, Butter, Bread} ⇒ {Egg}**

BASIC OF ASSOCIATION RULES:

- Frequent itemsets are used to obtain the association rules in the form of $X \Rightarrow Y$.

- Example of association rule: $\{Egg, Milk\} \Rightarrow \{Yogurt\}$.

- Based on this association rule, supermarket owners found that, commonly, customers who bought eggs and milk would also buy Yogurt

- Therefore, the supermarket can plan to promote yogurt to customers who often buy eggs and milk.

- Alternatively, the supermarket can arrange a shelf arrangement for yogurt sales near the egg and milk shelves.

BASIC OF ASSOCIATION RULES:

- The association rules describe the relationships or correlations between sets of items.

- Three basic measurements in choosing an association rules are:

i) Support

ii) Confidence

iii) Lift

- Support is the proportion of transactions in data that contain both item sets X and Y :

$$\text{support}\left(X \Rightarrow Y\right) = P\left(X \cap Y\right) = \frac{n_{XY}}{N}$$

- Confidence is the proportion of transactions that will contain item Y if item X has been purchased:

$$confidence(X \Rightarrow Y) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} = \frac{\left({n_{XY}}/{N} \right)}{\left({n_X}/{N} \right)}$$

- Lift is the ratio of Confidence to the proportion of transactions containing Y:

$$lift(X \Rightarrow Y) = \frac{confidence(X \Rightarrow Y)}{P(Y)} = \frac{P(X \cap Y)}{P(Y)P(X)} = \frac{\left({n_{XY}}/{N} \right)}{\left({n_X}/{N} \right)\left({n_Y}/{N} \right)}$$

- The higher the values of Support, Confidence and Lift, the higher the chance for item sets X and Y to occur together.

EXAMPLE:

- Given data for the following transactions data:

Transaction ID	Item Set
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Example of Association Rule: {Milk ,Diaper}⇒Beer

i) $$\text{support}\left(\{Milk, Diaper\} \Rightarrow Beer\right) = P\left(\{Milk, Diaper\} \cap Beer\right) = \frac{2}{5} = 0.4$$

ii) $$\text{confidence}\left(\{Milk, Diaper\} \Rightarrow Beer\right) = P\left(Beer \mid \{Milk, Diaper\}\right)$$

$$= \frac{P\left(\{Milk, Diaper\} \cap Beer\right)}{P\left(\{Milk, Diaper\}\right)} = \frac{\left({}^{2}\!/_{5}\right)}{\left({}^{3}\!/_{5}\right)} = \frac{2}{3} = 0.67$$

EXAMPLE:

- Given data for the following transactions data

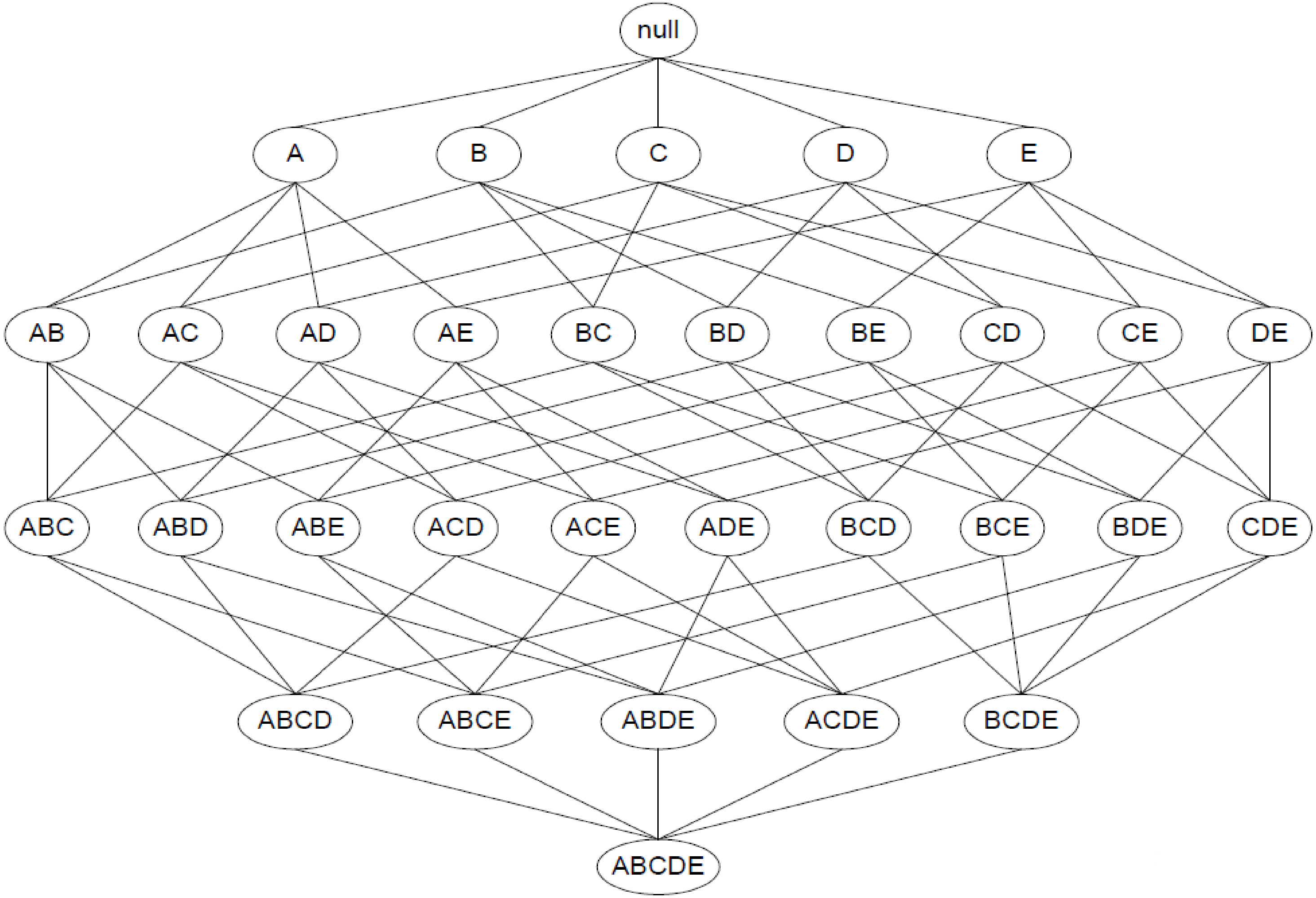
Transaction ID	Item Set
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Example of Association Rule: {Milk ,Diaper}⇒Beer

iii) $$lift\left(\{Milk, Diaper\} \Rightarrow Beer\right) = \frac{confidence\left(\{Milk, Diaper\} \Rightarrow Beer\right)}{P\left(Beer\right)} = \frac{0.67}{\left(3/5\right)} = 1.12$$

COMBINATIONS IN ALL ITEM SETS:

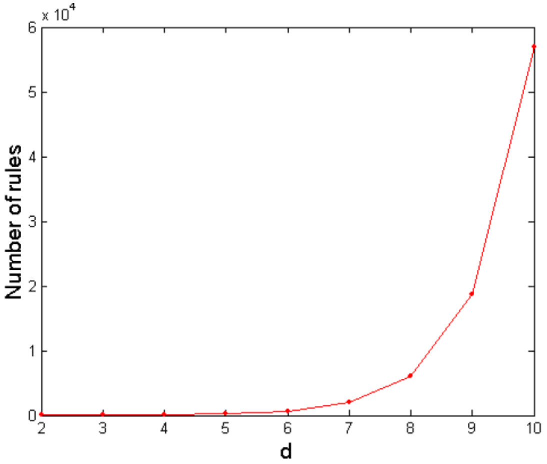
- If the supermarket has d=5 items, then there will be 2^d= 2^5= 32 possible sets of items that can be formed i.e.:



COMBINATIONS IN ALL ASSOCIATION RULES:

- Total combinations in all item sets = 2^d
- Then, the total of combinations in association rules:

$$R=\sum_{k=1}^{d-1}\left[\binom{d}{k}\times\sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$=3^{d}-2^{d+1}+1$$



- For example if *d* = 5, all possible association rules are 180.
- Generally, in actual data, the number of items *d* is very large.
- Then there are exist too many possible association rules.
- It is impossible to determine an association rules manually.
- A priori algorithms can be used to obtain the appropriate set of association rules.

ASSOCIATION RULES FRAMEWORK:

- All association rules $X \rightarrow Y$ must comply with the following framework:

$$\text{support}\left(X \cup Y\right) \geq \sigma$$

$$confidence\left(X \Rightarrow Y\right) \geq \gamma$$

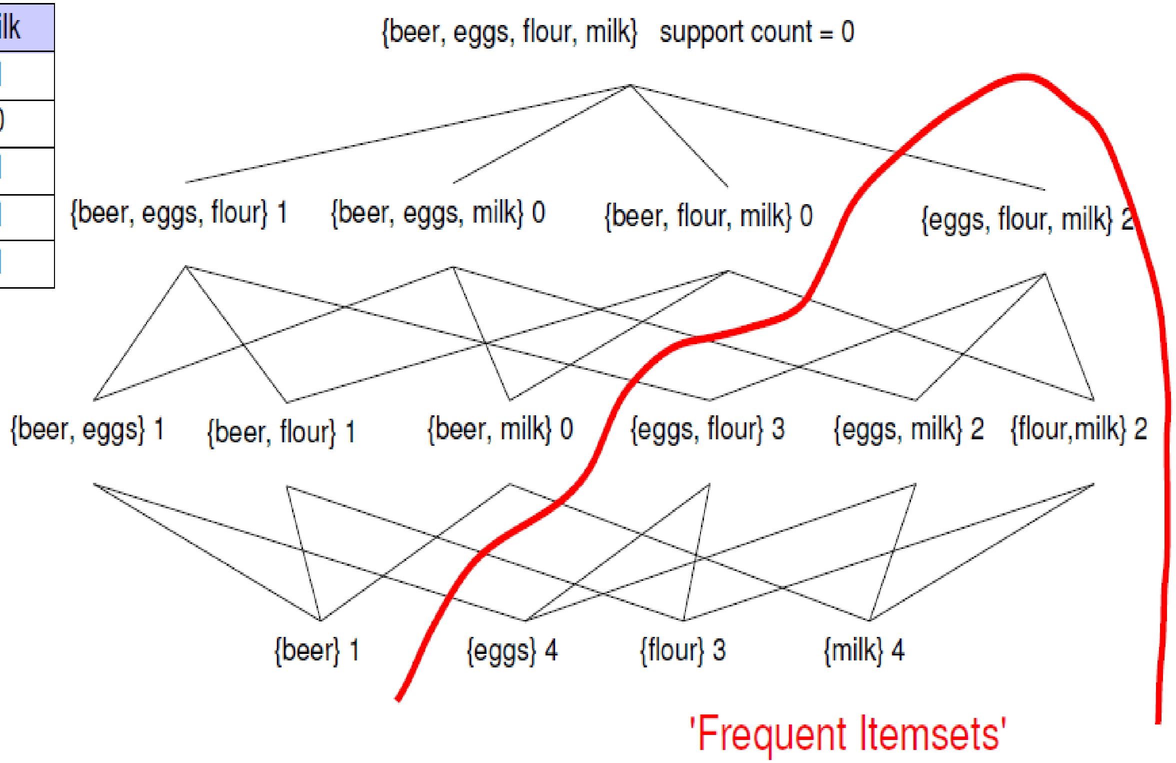
- σ is the minimum support threshold.
- γ is the minimum confidence threshold.

- The prerequisites for threshold values need to be determined by the analyst.
- Since there are too many combinations of association rules, thus a threshold values are very important for determining a meaningful association rules.

MINIMUM SUPPORT:

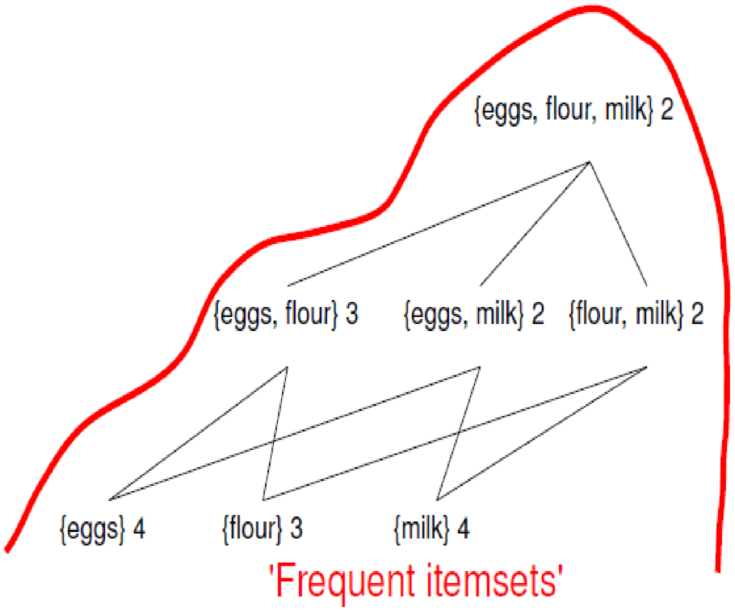
- Based on the minimum support value, only the most frequent itemset combinations will be retained.
- Example: For the following transaction data, with minimum support, σ=0.4:

Transaction ID	beer	eggs	flour	milk
1	0	1	1	1
2	1	1	1	0
3	0	1	0	1
4	0	1	1	1
5	0	0	0	1



MINIMUM CONFIDENCE:

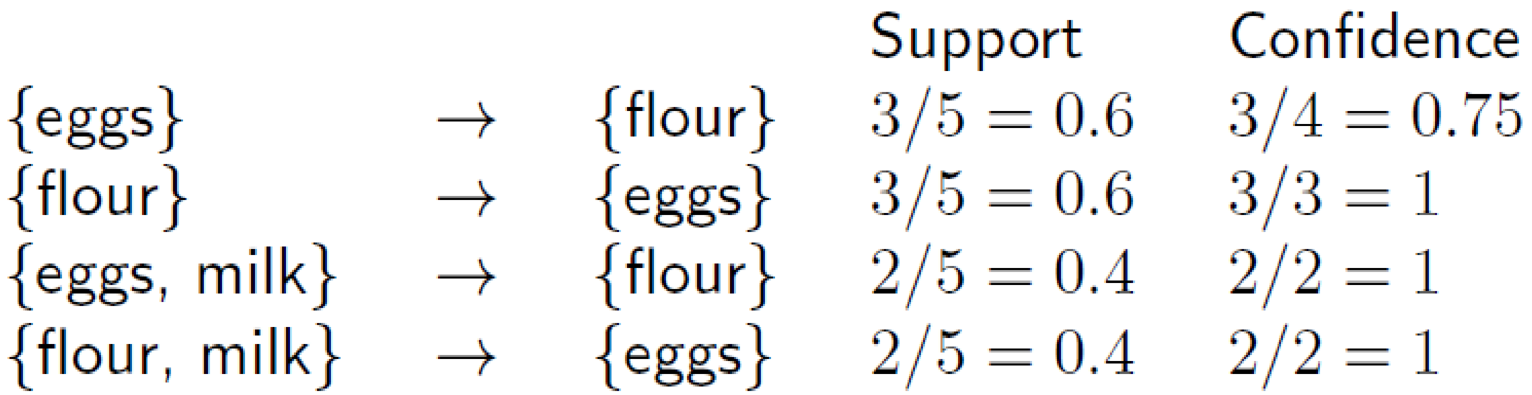
- For the following transaction data, with minimum confidence, γ =0.5.



			Confidence
{eggs}	→	{flour}	$3/4 = 0.75$
{flour}	→	{eggs}	$3/3 = 1$
{eggs}	→	{milk}	$2/4 = 0.5$
{milk}	→	{eggs}	$2/4 = 0.5$
{flour}	→	{milk}	$2/3 = 0.67$
{milk}	→	{flour}	$2/4 = 0.5$
{eggs, flour}	→	{milk}	$2/3 = 0.67$
{eggs, milk}	→	{flour}	$2/2 = 1$
{flour, milk}	→	{eggs}	$2/2 = 1$
{eggs}	→	{flour, milk}	$2/4 = 0.5$
{flour}	→	{eggs, milk}	$2/3 = 0.67$
{milk}	→	{eggs, flour}	$2/4 = 0.5$

LIFT MEASURE:

- Based on the minimum support and confidence, σ=0.5 and γ=0.7 set, only rule sets exceeding these prerequisites will be retained.



			Support	Confidence
{eggs}	→	{flour}	$3/5 = 0.6$	$3/4 = 0.75$
{flour}	→	{eggs}	$3/5 = 0.6$	$3/3 = 1$
{eggs, milk}	→	{flour}	$2/5 = 0.4$	$2/2 = 1$
{flour, milk}	→	{eggs}	$2/5 = 0.4$	$2/2 = 1$

- Next, the value of the lift can be calculated.

- The lift value obtained can be interpreted as :

i) $lift(X \Rightarrow Y) = 1$, X and Y is independent.

ii) $lift(X \Rightarrow Y) > 1$, X and Y has a complementary effect.

iii) $lift(X \Rightarrow Y) < 1$, X and Y has a substitute effect.

DETERMINATION OF ASSOCIATION RULES THROUGH ALGORITHMS APRIORI/ECLAT:

- In general, mining association rules can done through the following steps:

i) Determine all frequent item sets: Each set of items that occur more frequently (or equal to) than a predetermined minimum support threshold will be identified.

ii) Determine the association rules from (i): Each set of items that meets the minimum support threshold property also meets the minimum confidence threshold property.

APPLICATION IN R:

- In R, mining association rules can be done through a priori or eclat algorithm.

- Among the important things that need to be determined are:

i) How to determine the most frequent items?

ii) How to obtain association rules for product recommendations?

iii) How to remove redundant rules?

iv) How to determine the association rules related to some particular item?

REFERENCES:

- Adamo, J-M. (2001). *Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms*. Springer.
- Aggarwal, C.C. (2015). *Data Mining: The Textbook*. New York: Springer.
- Hahsler, M., Grün, B., Hornik. K. (2005). arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, 14(15), 1–25.
- Makhabel, B. (2015). *Learning Data Mining with R: Develop key skills and techniques with R to create and customize data mining algorithms*. Birmingham: Packt Publishing.
- Subramanian, G. (2017). *R Data Analysis Projects: Build end to end analytics systems to get deeper insights from your data*. Birmingham: Packt Publishing.

NEXT TOPIC:

Mining Time Series Data