Association Analysis: Introduction

James Kwok

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Motivating Example

- the table below depicts a set of transactions at a grocery store

Market basket transactions example

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- each transaction corresponds to the set of items purchased by a single customer at the checkout counter of the store
- such transactions are also called market basket transactions

Which items are purchased together by the customers?

- given a set of transactions, association analysis finds rules that predict the occurrence of an item in a transaction based on the occurrences of other items in this transaction

Market basket transactions example

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- one rule that can be extracted is that diapers $\rightarrow$ beer
  - this rule suggests a strong relationship between the sale of diapers and beer: customers who buy diapers also buy beer
- retailers can use this information for cross-marketing, catalog design, customer shopping behavior analysis, ...

Other applications

- medical diagnosis, DNA sequence analysis, web log (clickstream) analysis, ...

Gender	Age	Smoking	Blood pressure	...	Class
M	40 - 50	Y	high	...	abnormal
M	20 - 40	N	normal	...	normal
F	20 - 40	N	normal	...	normal
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$

Binary Representation

- transactional data can be stored in binary format
- the binary representation of the data in the example of the previous slides is depicted below

TID	Bread	Milk	Diaper	Beer	Eggs	Coke
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

- a '1' entry implies the presence of the item in the transaction, whereas a '0' entry implies its absence

Support

Market basket transactions example

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- $I = \{I_1, I_2, \ldots, I_n\}$: set of all items appearing in the transactional database
- itemset: a set of items (that is a subset of I)
- support count of an itemset A (support_count(A))
  - number of transactions that contain A
  - example: support_count($\{\texttt{Milk, Diaper, Beer}\}$) $= 2$
- support of an itemset A
  - fraction of transactions that contain A
  - let $|T|$ denote the number of transactions in the database
  - support(A) = support_count(A)$/|T|$
    - alternatively, the probability that a transaction contains A: support(A) $= P(A)$
  - example: support($\{\texttt{Milk, Diaper, Beer}\}$) $= 2/5$

Association Rule

- let $A \subset I$ and $B \subset I$ be two itemsets, such that $A \cap B = \emptyset$

Example

$A = \{\texttt{Milk, Diaper}\}$ and $B = \{\texttt{Beer}\}$

- association rule: implication of the form
  $A \to B$ [*support*, *confidence*]

Example

$\{\texttt{Milk, Diaper}\} \to \{\texttt{Beer}\}$ [*support*=40%, *confidence*=67%]

Support of a Rule $A \rightarrow B$

Market basket transactions example

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- fraction of transactions that contain both A and B
- $\text{support}(A \rightarrow B) = \text{support\_count}(A \cup B)/|T|$
- example: $\text{support}(\{\texttt{Milk, Diaper}\} \rightarrow \{\texttt{Beer}\}) = 2/5 = 40\%$
- alternatively, probability that a transaction contains both A and B
  - $\text{support}(A \rightarrow B) = P(A \cup B)$
  - note: $P(A \cup B)$ is NOT probability of A or B

Why use support?

low support rules

- may occur simply by chance
- may not be interesting

Confidence of a Rule $A \to B$

Market basket transactions example

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- fraction of transactions containing A that also contain B

$$\text{confidence}(A \to B) = \frac{\text{support}(A \cup B)}{\text{support}(A)} = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)}$$

- example: confidence($\{\texttt{Milk, Diaper}\} \to \{\texttt{Beer}\}) = 2/3 = 67\%$
- alternatively, conditional probability that a transaction having A also contains B
  - confidence$(A \to B) = P(B|A)$

Why use confidence?

- measures the reliability of the inference made by a rule

Strong Rules

- define two parameters
  - minimum support threshold (or minimum support count threshold) *min_sup*
  - minimum confidence threshold *min_conf*
- a strong association rule satisfies both *min_sup* and *min_conf*

How to mine the strong association rules?

Brute-force

- compute the support and confidence for every possible rule
- computationally expensive

Idea

Decouple the support and confidence requirements

- support of a rule $X \rightarrow Y$ depends on only the support of its corresponding itemset $X \cup Y$

Frequent itemset

- itemset A is frequent if it satisfies *min_sup*
- A is a frequent k-itemset if it is frequent and contains k items

Suppose that we have already computed all frequent itemsets

- for each frequent itemset S, generate all nonempty proper subsets A of S
- for each A, output rule $A \rightarrow (S - A)$ if
$$\text{confidence}(A \rightarrow (S - A)) = \frac{\text{support\_count}(S)}{\text{support\_count}(A)} \geq \mathit{min\_conf}$$
- this rule automatically satisfies *min_sup* since it is derived from a frequent itemset
- hence, a strong rule

Example

Market basket transactions example

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- let $\{\texttt{Diaper, Beer}\}$ be a frequent itemset
- its proper nonempty subsets are $\{\texttt{Diaper}\}$ and $\{\texttt{Beer}\}$
- the rules are

$$\{\texttt{Diaper}\} \rightarrow \{\texttt{Beer}\} \text{ with confidence } 3/4 = 75\%$$
$$\{\texttt{Beer}\} \rightarrow \{\texttt{Diaper}\} \text{ with confidence } 3/3 = 100\%$$

- if $min\_conf = 90\%$, only the second rule is strong

Another Example

Example

- frequent itemset: $\{i1, i2, i5\}$
- subsets: $\{i1, i2\}$, $\{i1, i5\}$, $\{i2, i5\}$, $\{i1\}$, $\{i2\}$, $\{i5\}$
- resultant association rules
  - $\{i1, i2\} \rightarrow i5$, confidence $= 2/4 = 50\%$
  - $\{i1, i5\} \rightarrow i2$, confidence $= 2/2 = 100\%$
  - $\{i2, i5\} \rightarrow i1$, confidence $= 2/2 = 100\%$
  - $i1 \rightarrow \{i2, i5\}$, confidence $= 2/6 = 33\%$
  - $i2 \rightarrow \{i1, i5\}$, confidence $= 2/7 = 29\%$
  - $i5 \rightarrow \{i1, i2\}$, confidence $= 2/2 = 100\%$
- if the minimum confidence threshold is 70%, then only the second, third and last rules are strong

Association Rule Mining

1. find all frequent itemsets
   - by definition, all these itemsets satisfy *min_sup*
2. generate strong association rules
   - analyze the frequent itemsets further to extract rules that also satisfy *min_conf*

Strong Rules are Not Necessarily Interesting

Example

- suppose we have 10,000 transactions
  - 6,000 include $\{\texttt{Games}\}$
  - 7,500 include $\{\texttt{Videos}\}$
  - 4,000 include $\{\texttt{Games, Videos}\}$

  $\{\texttt{Games}\} \rightarrow \{\texttt{Videos}\}$ [*support*= 40%, *confidence*= 66%]
- suppose *min_sup*=30% and *min_conf*=60%, is this rule strong? yes
- is this rule useful?
  - the probability one buys videos is $\frac{7{,}500}{10{,}000}$ =75% $>$ 66%
  - knowing that one buys games actually decreases her probability of buying videos
- one could take unwise business decisions based on the above rule