TIM 245 Lecture 17 (5/31/17)

Agenda

1) Review Lecture 16

2) Association Analysis Basic Concepts

3) Apriori Algorithm

4) Evaluation of association rules

① Association Analysis Basic Concepts

<u>objective</u> : given a set of transactions, predict the occurence of an item based on the occurence of other items in the transaction

Transactions

ID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beers, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, coke

Association Rules

{ Diapers } → { Beer }

{ Bread } → { Milk }

{ Beer, Bread } → { Milk }

Transactions can be any set of "Items" that occured together:

- Netflix user : movies
- Text : words
- Students : courses
- Papers : Co authors
- Patients : diseases
- Gamers : in-app purchases

Itemset: a collection of one or more items

e.g. $X = \{milk, diapers\}$

Support Count (σ): frequency of occurence of an itemset

e.g. $\sigma(X) = 3$

Support: fraction of transactions that contain a itemset

e.g. $Support(X) = \frac{\sigma(X)}{n} = \frac{3}{5} = 0.6$

Frequent Itemset: itemsets with support greater than or equal to minsup, a user provided threshold

Association Rule: implies the co-occurence of itemsets X and Y

e.g. $X \rightarrow Y$

where $Y = \{beer\}$

Association rules can be evaluated based on two metrics:

Support (s): fraction of transactions that contain both X and Y (impact)

e.g. $s(X \rightarrow Y) = \frac{\sigma(\text{Milk, Diapers, Beer})}{n} = \frac{2}{5}$

Confidence (c): how often items in Y appear in transactions that contain X (usefulness)

e.g. $c(X \rightarrow) = \frac{\sigma(\text{Milk, Diapers, Beer})}{\sigma(\text{Milk, Diapers})} = \frac{2}{3}$

How do we find all the rules that are above a minsup and minconf threshold?

Bruteforce approach: enumerate all possible association rules and prune rules below minsup and minconf threshold.

Rules originating from the same itemset also have the same support. Therefore we can use a two-step approach.

1) Generate all the itemsets whose support greater than or equal to minsup
(Frequent itemset generation)

2) Generate high confidence rules from each frequent itemset where each rule is a binary partition of the frequent itemset
(rule generation)

How do we efficiently generate frequent itemsets?

③ Apriori Algorithm

Basic idea: reduce the number of candidate itemsets through early pruning

Apriori principle: if an itemset is frequent then all of its subsets must also be frequent.

$$\forall X, Y : X \subseteq Y \rightarrow s(X) \geq s(Y)$$

Support of an itemset never exceeds the support of its subsets. Conversely, if an itemset is infrequent then all all of its supersets must be infrequent too.

We can use this property for early pruning during frequent itemset generation.

Apriori Algorithm

1) Initialize K to 1

2) Generate frequent itemsets of length 1

3) Generate length K+1 candidate itemsets from length K frequent itemsets

4) Compute support for candidates and prune infrequent itemsets below minsup

5) Increment K and repeat until no new frequent itemsets are identified.

(4) Evaluation of Association Rules

Given a rule $X \rightarrow Y$ we want to know how interesting it is.

High confidence rules can be deceiving because they don't take the support of Y into account.

Example: Tea $\rightarrow$ Coffee is a high confidence rule, e.g. 80%. However 80% of the population drinks Coffee.

A variety different statistical measures for pre-screening interestingness can be computed from the contingency table.

	Y	$\bar{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\bar{X}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

f_{11} : Support X and Y

f_{00} : Support $\bar{X}$ and $\bar{Y}$

Most Common measure of interestingness is the lift ratio.

$$\text{Lift}(X \to Y) = \frac{c(X \to Y)}{s(Y)} \quad \frac{(\text{actual})}{(\text{expected})}$$

which can be computed from the contingency table as

$$\text{Lift}(X \to Y) = \frac{s(X, Y)}{s(X) \cdot s(Y)} = \frac{N f_{11}}{f_{1+} \cdot f_{+1}}$$

Values above 1 means that X and Y are positively correlated and that the rule has some objective measure of interestingness.

Many other measures can be computed from the contingency table

- Gini index
- Correlation
- Jaccard
- ....

Ultimately rules need to be evaluated by a domain expert for interestingness