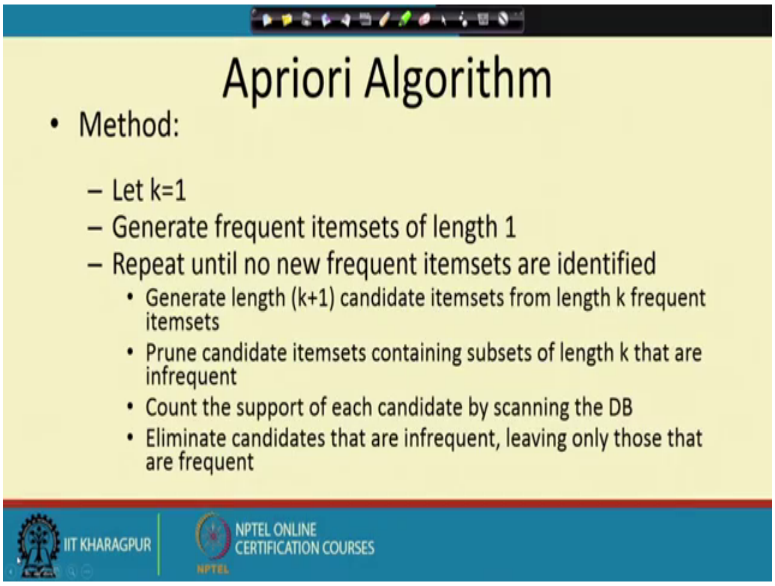
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 06
Rule Generation

(Refer Slide Time: 00:28)

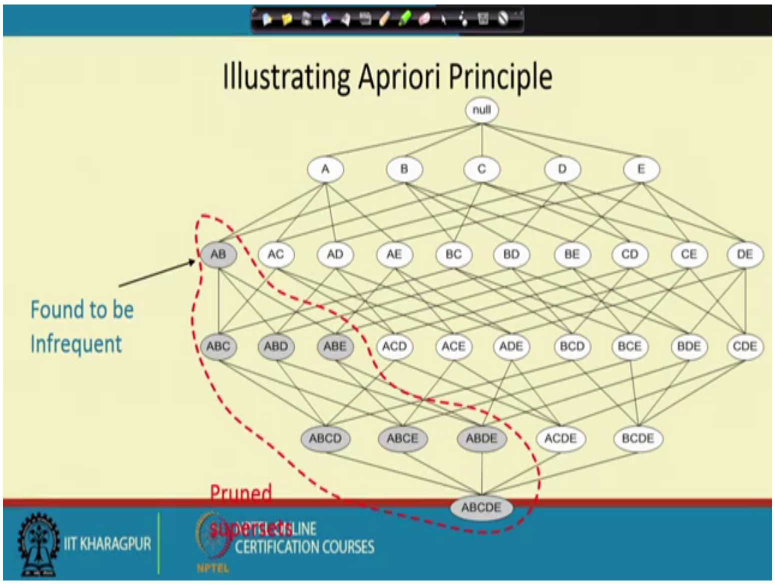


Apriori Algorithm

- Method:
  - Let k=1
  - Generate frequent itemsets of length 1
  - Repeat until no new frequent itemsets are identified
    - Generate length (k+1) candidate itemsets from length k frequent itemsets
    - Prune candidate itemsets containing subsets of length k that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

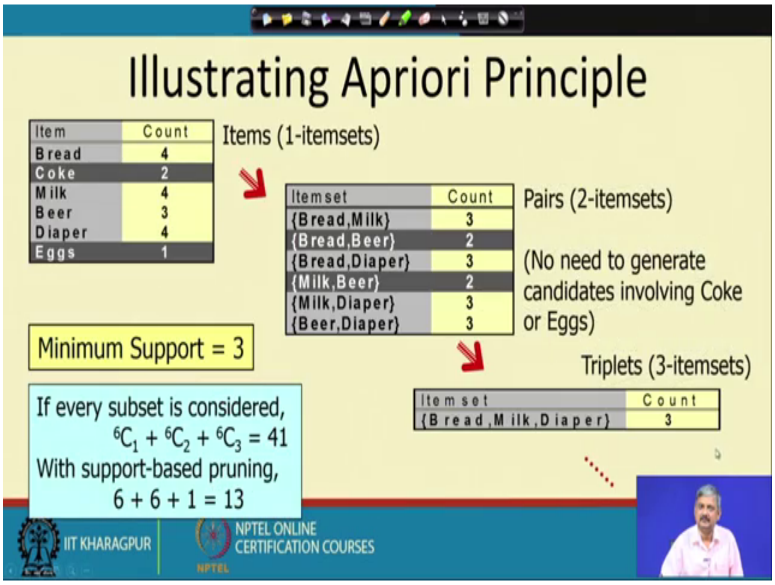
Welcome to the third lecture on association rule binding. We have seen the apriori algorithm. What I do in apriori algorithm is to reduce the number of candidate item sets using the apriori principal. So, if you recollect the steps of finding the association rule, so what we do is that first we suppose there are m number of or d number of items, we will try all possible item sets in X and y. And see if their combination x, Y is frequent that means, they appear more than means of fraction of total transactions. And once we identify the frequent item sets next step will do the rule generation.

(Refer Slide Time: 01:21)


Illustrating Apriori Principle

So, if you recollect that the all possible item sets, for example if we have five items A, B, C, D, E can be arranged in a lattice like this. And if we do not do any intelligent thing what you have to do is that check whether each of this combinations, how many times they appear in the total number of transactions, and then threshold them based on the means up and tell them to be. What the a priori principle says is that if some item set AB is not frequent then none of your super sets ABC, ABD, ABCD, ABC can be frequent. So, for example, if milk and butter themselves are not frequent, milk butter sugar three cannot be frequent. If two are not frequent, three cannot be frequent; there obviously, three will appear in less number of transactions than just two. So, this principle we use to prune them we just can remove among the candidates and evaluate for the rest.

(Refer Slide Time: 02:44)



Illustrating Apriori Principle

Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Pairs (2-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

(No need to generate candidates involving Coke or Eggs)

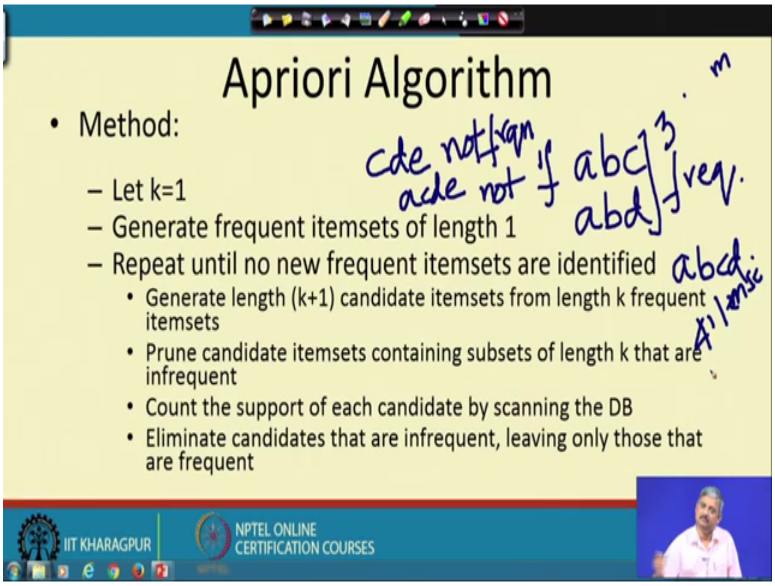
Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

Minimum Support = 3

If every subset is considered,
$^6C_1 + {^6C_2} + {^6C_3} = 41$
With support-based pruning,
$6 + 6 + 1 = 13$

(Refer Slide Time: 02:46)



Apriori Algorithm

- Method:
  - Let k=1
  - Generate frequent itemsets of length 1
  - Repeat until no new frequent itemsets are identified
    - Generate length (k+1) candidate itemsets from length k frequent itemsets
    - Prune candidate itemsets containing subsets of length k that are infrequent
    - Count the support of each candidate by scanning the DB
    - Eliminate candidates that are infrequent, leaving only those that are frequent

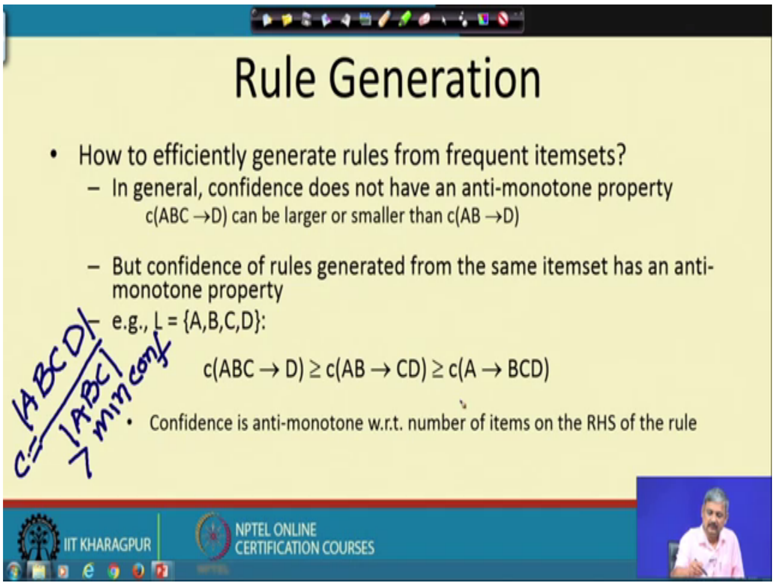
m
cde not freq
acde not f
if abc
abd
3 freq.
abcd
4 itemsets

So, this was used in the apriori algorithm. What we do is that go down this way from the lattice, we go down this way from the lattice. Take the single item sets, which are frequent. Then from the single item sets combine to produce frequent two item sets. And from the two, produce three item sets. So, how do you produce two from three? You see that suppose AE and AD, they are frequent you can join these two if they differ by or if they have if they differ by only one. So, here D and E are differed. So, if we join them you get ADE. So, A, D and E, if we join we take we get ADE.

So, what we do in the apriori principal is that from k frequent item sets, we produce k plus 1 sized frequent item sets, how by joining two items which are sharing some of the items except one. So, ABC and ABD we these are three item sets delta they will join to produce a four item set ABCD, ABC, ABD joins to produce ABD. So, let me write down. So, if I have if abc is frequent and abd is frequent, I can join them they are frequent I can join them to produce abcd as candidate frequent. Note that it may be frequent it may not be frequent. So, to actually take if it is frequent I have to again go through this transactions, but if abc is not frequent suppose cde is not frequent not frequent then definitely acd is also not frequent not frequent if something is not frequent its superset cannot be frequent.

So, this way from three item sets I produce four item sets; from three item sets, I produce four item sets of size three and size four. So, this way you go on doing and checking. Now, it may turn out that abcd itself is not frequent abcd itself abc are frequent abd are frequent abcd is not frequent then definitely abcde cannot be frequent, this way we go on.

(Refer Slide Time: 06:09)



Rule Generation

- How to efficiently generate rules from frequent itemsets?
  - In general, confidence does not have an anti-monotone property
    c(ABC →D) can be larger or smaller than c(AB →D)
  - But confidence of rules generated from the same itemset has an anti-monotone property
  - e.g., L = {A,B,C,D}:

    $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$

    - Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

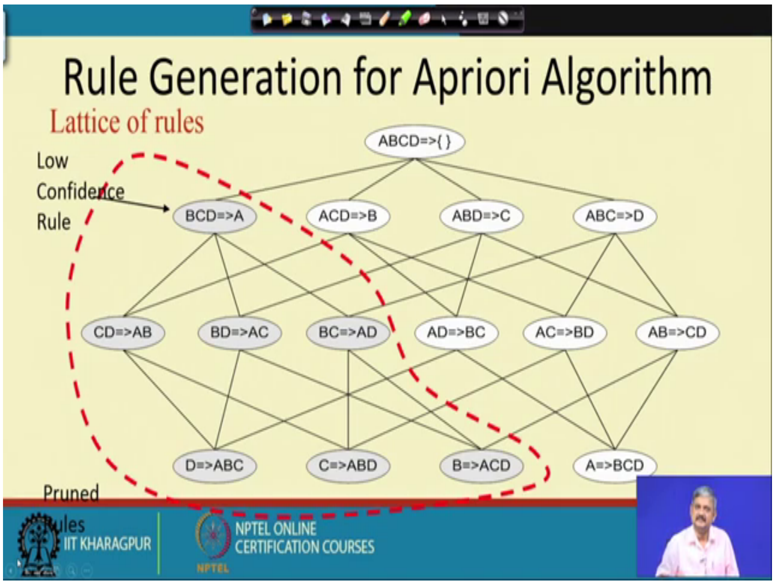
$c = \frac{|ABCD|}{|ABC|} > min\ conf$

Now, let us come to the next part about rule generation. So, we now have some item sets which are known to be frequent. What we want to do is that now split them into left and right halves. So, I know say ABCD is frequent say for example, ABCD is frequent. So, what are the rules I can produce form ABCD, I can produce ABC associated with D, I can produce AB associated to it CD, I can produce A associated with BCD. So, now, using this frequent item set I test by the (Refer Time: 06:57) but now whether testing whether the rules are frequent I need to test no longer the support criteria because support they are already highly supported, they are frequent item sets. I need to test the confidence criteria.

What is the confidence criteria, that means, the fraction of time only D appears the ratio of number of the times only D appears or rather the ratio of number of times ABCD appears to the ratio of the number of times only ABC appear that is the confidence. So, it is number of times sorry number of times ABCD appears total number of times divided by number of times ABC appears that is my confidence. If this is greater than some threshold called the min conf I would say it is a valid rule. Now, given a single item set I can have various such partitions. So, should I check all of them the thing is that even here there is something like that apriori or anti-monotone property.

So, what it says is that if you take the same set of items, let us say ABCD then I have the rule that if the number of elements in the left hand side are more than its confidence is more than this. You can naturally see from here that more number of elements you have in the left hand side, so the more that denominator side size is the larger this ratio will be more the denominator side is larger this ratio will be. Because more and more things will be included, because ABCD appears in much less fraction of only A or only AB as opposed to ABC. So, you can see here this ordering is in terms of the size of the left hand side ABC, AB, A even in goes to the right hand side. Their confidence can be ordered like this. This you can easily verify, this follows directly from the definition of confidence. You can take a minute and verify this that this will always be true. If this is true, if this is true, I can now form a lattice among the rules like this.

(Refer Slide Time: 10:19)



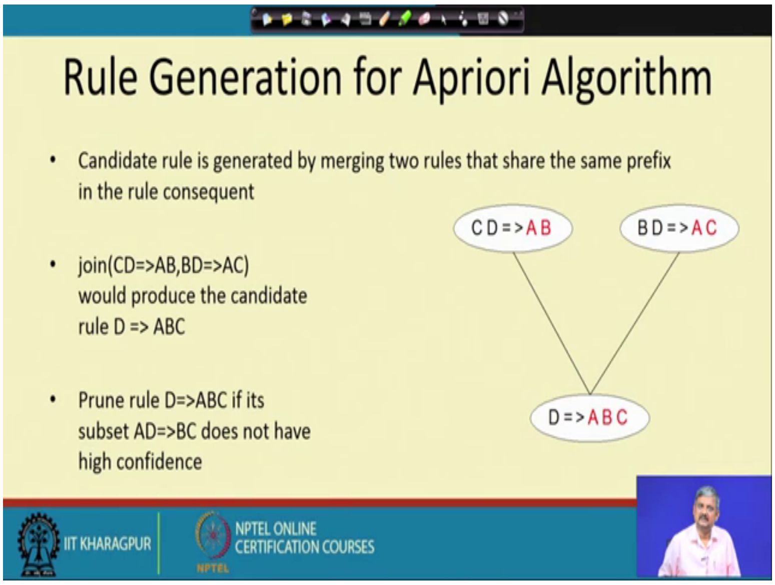
Rule Generation for Apriori Algorithm

Lattice of rules

How do I form a lattice among the rules in the top of the lattice I take all the attributes in the left hand side and then gradually move one attribute to the right and then two attribute to the right and so on. So, what is the relation here is that the left hand side CD is a subset of BCD; similarly the left hand side BD is a subset of BCD that way you will form like. D is a subset of CD, D is a subset of BD, D is a subset of AD, we look at so this relations lattice and these edges are basically the checking for subsetness of the left hand side.

So, now, you can actually check that now there is a ordering. So, these more number of items in the left hand side lower the confidence. We have already talked about that ordering in the previous slide you have this ordering. So, these will have low confidence than the others. So, you can see that it follows directly from that rule that if this is found to be below the confidence threshold, all these rules will also be below the confidence threshold; in other words, you can prune the lattice.

(Refer Slide Time: 12:26)

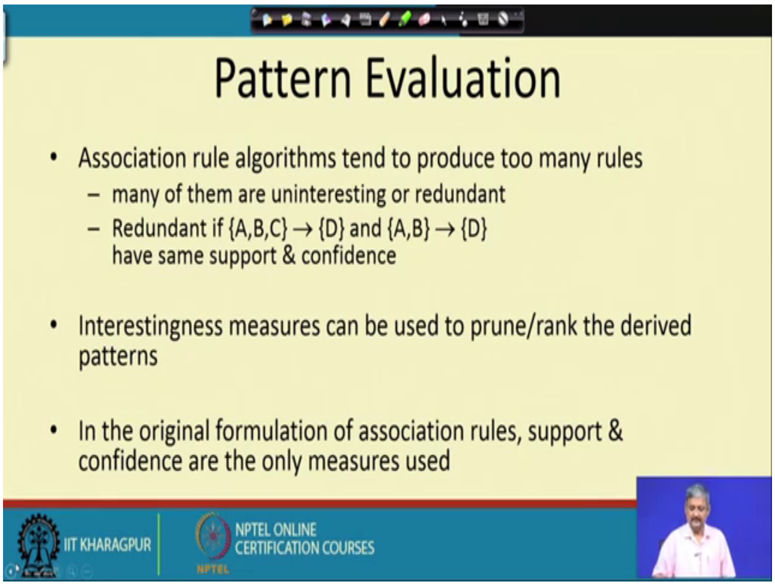


Rule Generation for Apriori Algorithm

- Candidate rule is generated by merging two rules that share the same prefix in the rule consequent
- join(CD=>AB,BD=>AC) would produce the candidate rule D => ABC
- Prune rule D=>ABC if its subset AD=>BC does not have high confidence

So, this gives me an principle of obtaining the rules suppose CD associated it AB, and BD associated AD AC are two valid rules. We can join them, if the right hand side have the same prefix. So, we can join them by moving one rule from here the common rule from here to the other. So, ABC I can move. So, this will be high confidence.

(Refer Slide Time: 13:18)

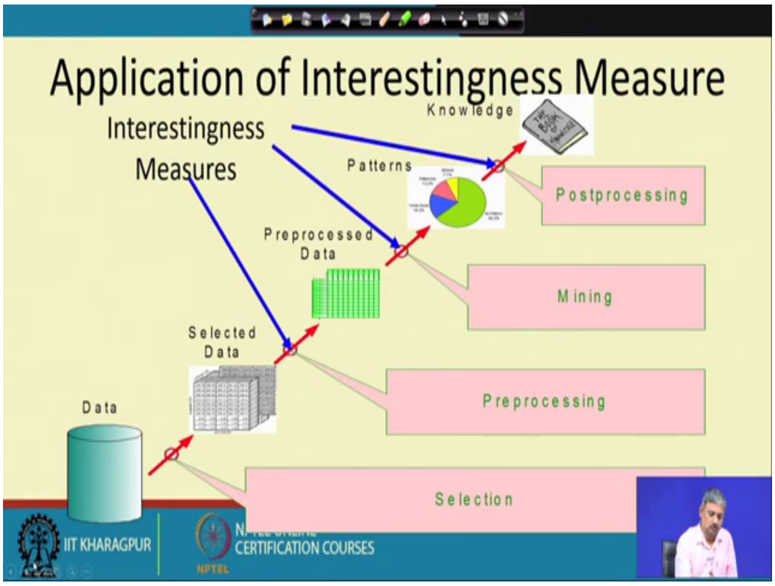


Pattern Evaluation

- Association rule algorithms tend to produce too many rules
  - many of them are uninteresting or redundant
  - Redundant if {A,B,C} → {D} and {A,B} → {D} have same support & confidence
- Interestingness measures can be used to prune/rank the derived patterns
- In the original formulation of association rules, support & confidence are the only measures used

So, this is how you get the rules. The thing is that after all this process you get a number of such rules. And they satisfy the support and confidence. The thing is that sometimes even after satisfying support and confidence criteria you need to evaluate using a third criteria and further criteria whether these are valid rules.

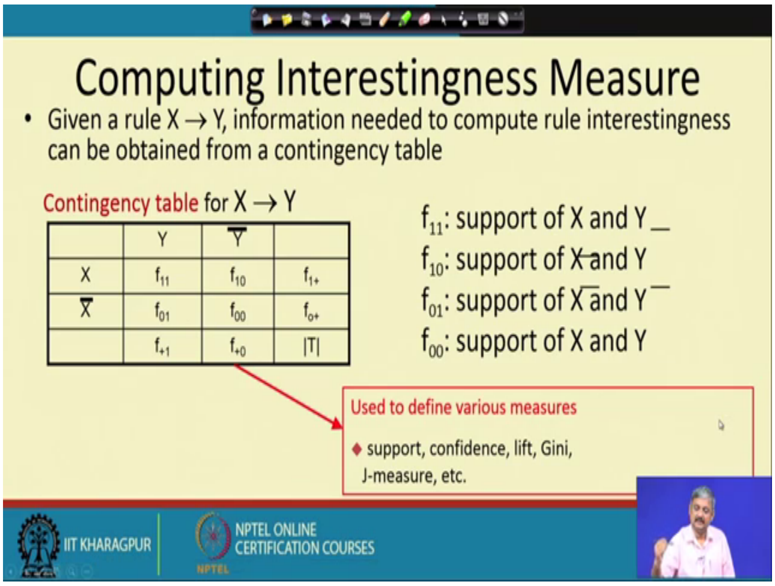
(Refer Slide Time: 13:54)



Application of Interestingness Measure

So, this was actually done in the general process of data mining itself you use something called a interestingness measure of a rule how interesting a rule is you apply it in several places preprocessing mining to get it.

(Refer Slide Time: 14:13)



Computing Interestingness Measure

- Given a rule X → Y, information needed to compute rule interestingness can be obtained from a contingency table

Contingency table for X → Y

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	|T|

f_{11}: support of X and Y

f_{10}: support of X and $\overline{Y}$

f_{01}: support of $\overline{X}$ and Y

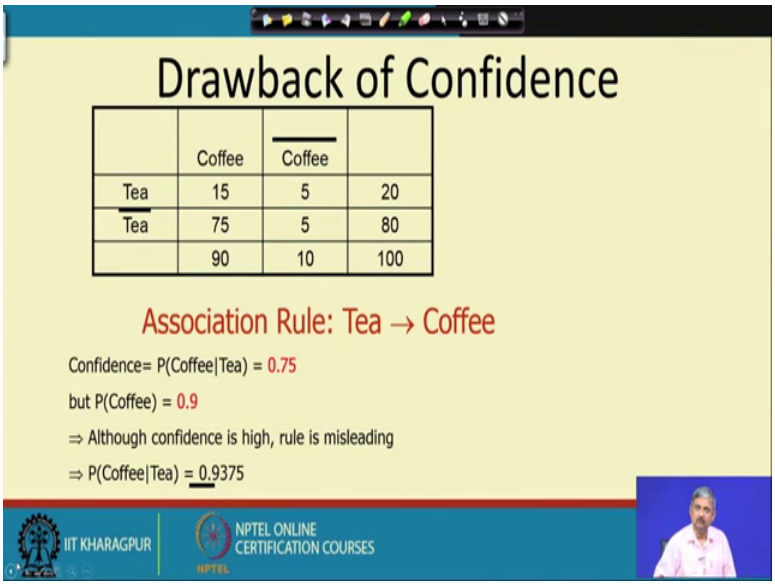
f_{00}: support of $\overline{X}$ and $\overline{Y}$

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

So, we will now briefly discuss for the rules extracted by the previous algorithm. What are the evaluations in terms of the interestingness? So, for that I use a contingency table which means like this. Suppose, I have a rule X to Y discounts a number of basically supports of how many times X happens as well as Y happens, X happens as well as not Y happens, not X happen as well as Y happens and so on. So, this is the definition. So, pardon me for this mistake. These bars will come to the top they mean non none not this X and Y. Now you can define various measures on this I will mention about some of them.

(Refer Slide Time: 15:12)



Drawback of Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence= P(Coffee|Tea) = 0.75

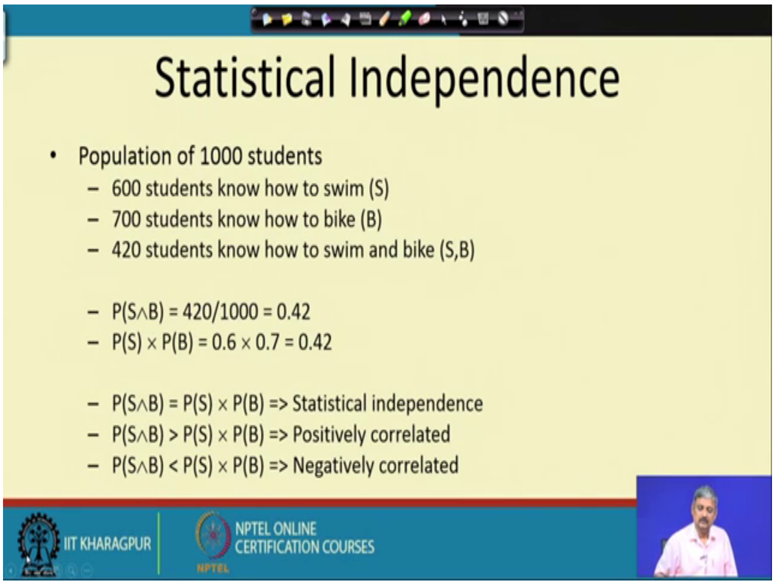
but P(Coffee) = 0.9

⇒ Although confidence is high, rule is misleading

⇒ P(Coffee|Tea) = 0.9375

So, let me give a motivating example suppose this is a contingency table and I association rule like this and these are my confidence factor. So, even though if you have a high confidence factor, this rule may be misleading that people who buy tea also buy coffee.

(Refer Slide Time: 15:36)



Statistical Independence

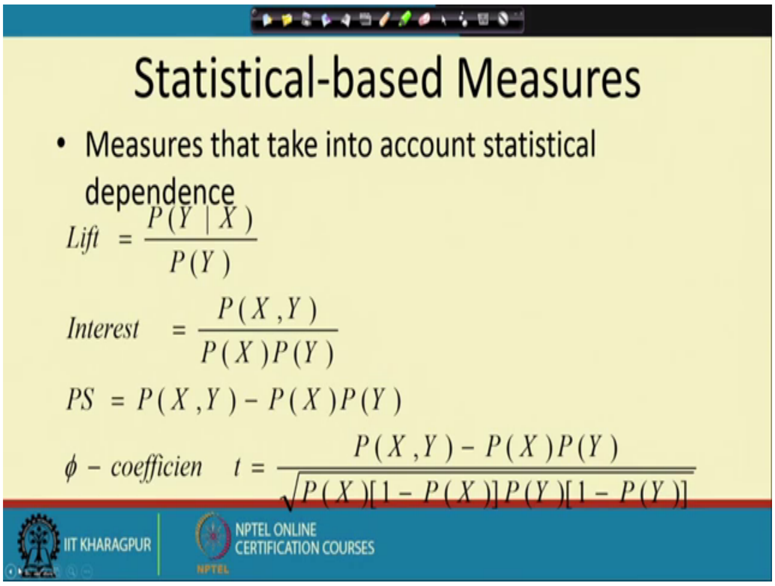
- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S,B)
  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
  - $P(S \wedge B) = P(S) \times P(B)$ => Statistical independence
  - $P(S \wedge B) > P(S) \times P(B)$ => Positively correlated
  - $P(S \wedge B) < P(S) \times P(B)$ => Negatively correlated

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, you need some additional measures. The root of all these methods a principle called statistical independence. So, here in this case the events are people buying some S and people buying some B. So, here is an something else example. So, students swimming or student biking. So, swimming and biking. So, we have the joint probability both S and B, and the product probability of individual S and individual B, if they are exactly same we say they are independent; if this is greater than this we say they are positively correlated; if it is less we say they are negatively correlated.

(Refer Slide Time: 16:23)



Statistical-based Measures

- Measures that take into account statistical dependence

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

$$PS = P(X,Y) - P(X)P(Y)$$

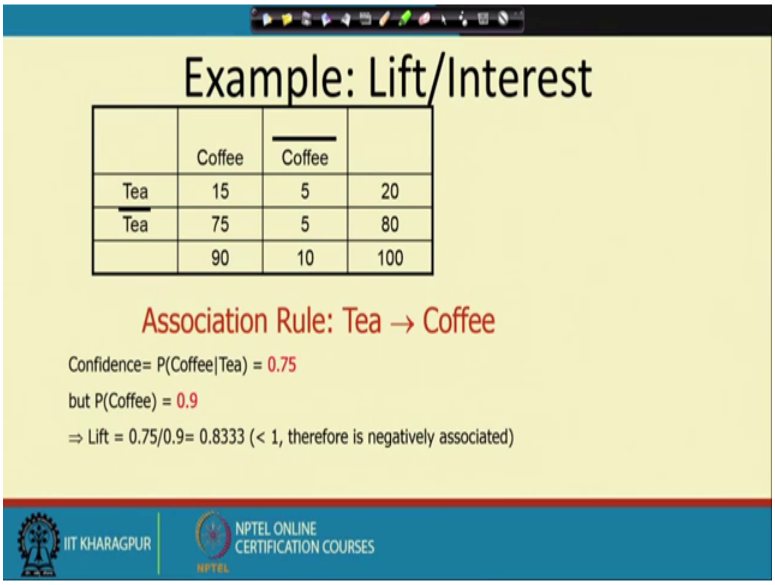
$$\phi - coefficien \quad t = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, using these probabilities one can define measures like this, lift, interest, PS some PS value phi coefficient. So, they are Y given X divided by only Y is lift this is a very common measure used. So, we will see how to define these terms in terms of the contingency table, but first probabilistically define them. So, you have a lift which is Y given X by Y, you have interest given by joint probability by product of the individual probabilities which is nothing but the this ratio of this quantity to this quantity. You have PS, which is difference of P X Y and minus P X minus P Y. Note that each of this quantity interest it would be one if they are independent; PS would be 0, if they are independent lift; PS would be 1, if they are independent.

(Refer Slide Time: 17:48)



Example: Lift/Interest

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence= P(Coffee|Tea) = 0.75

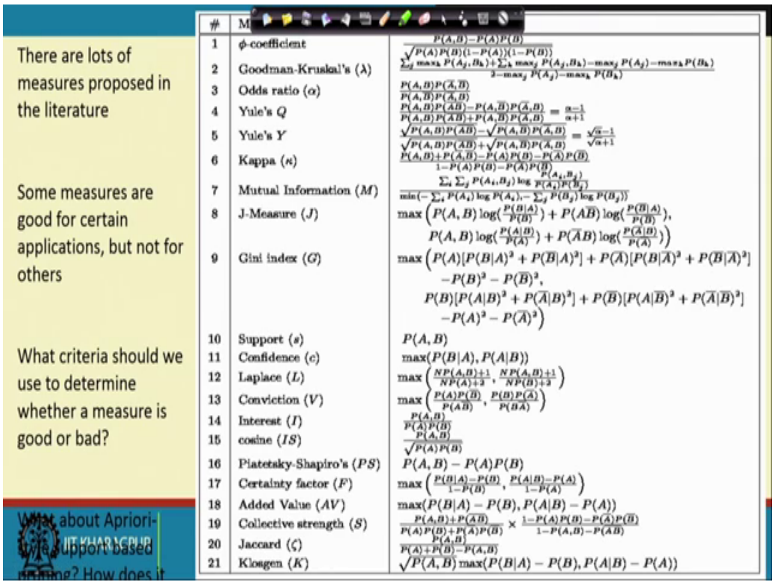
but P(Coffee) = 0.9

⇒ Lift = 0.75/0.9= 0.8333 (< 1, therefore is negatively associated)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if we have this kind of contingency table, if you remember the definition of it means 14 items, 14 transactions are both tea and coffee, five transactions have tea does not have coffee, 75 transactions does not have tea have coffee and so on. So, the lift is 0.9. So, it is negatively if the lift is my less than 1, it is negatively associated. So, even though confidence is high they are actually negatively associated.

(Refer Slide Time: 18:25)



There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

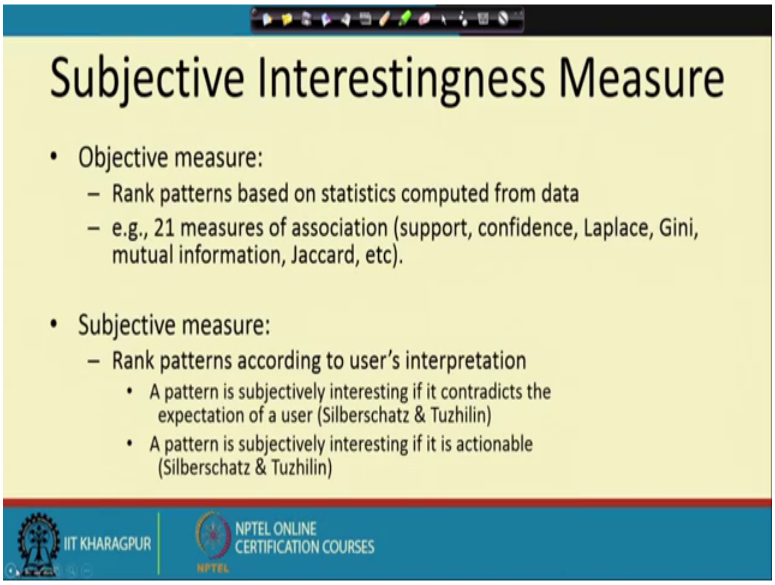
What criteria should we use to determine whether a measure is good or bad?

What about Apriori-style support based pruning? How does it

#	M	
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A})[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2]$ $-P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B})[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2]$ $-P(A)^2 - P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B) - P(A)P(B)$
17	Certainty factor (F)	$\max\Big(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\|A) - P(B), P(A\|B) - P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$

So, here is a list of measures and that one might use I have talked about the psi coefficient and the PS measure (Refer Time: 18:48) measure, there are Jaccard coefficients. So, A and B are the item sets. So, probability you can read at what fraction of the total item sets these A and B appears. So, support is simply both A and B appear for the rule A to B for the rule A to B, both A and B appear. So, I can define all these different quantities as my interestingness measures, you can have a look at them I can use all these quantities as my interestingness measures. They are variations of each other in certain situation certain things work well other situations other thing work well.

(Refer Slide Time: 19:42)



Subjective Interestingness Measure

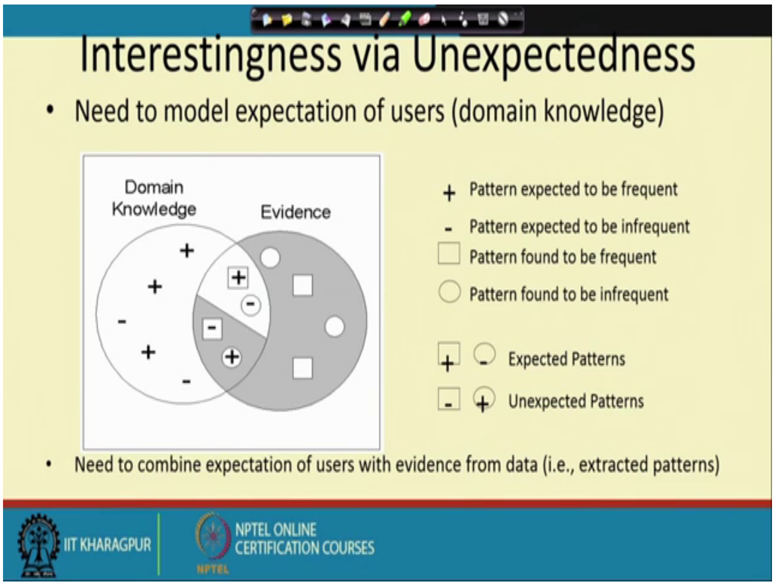
- Objective measure:
  - Rank patterns based on statistics computed from data
  - e.g., 21 measures of association (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).
- Subjective measure:
  - Rank patterns according to user's interpretation
    - A pattern is subjectively interesting if it contradicts the expectation of a user (Silberschatz & Tuzhilin)
    - A pattern is subjectively interesting if it is actionable (Silberschatz & Tuzhilin)

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

Similarly, there are other interesting measures, which are subjective based on the domain, based on the domain of the problem, so you can define. All this once I talked about are all objective that is independent of domain.

(Refer Slide Time: 20:08)



Interestingness via Unexpectedness

- Need to model expectation of users (domain knowledge)

+ Pattern expected to be frequent

- Pattern expected to be infrequent

□ Pattern found to be frequent

○ Pattern found to be infrequent

⊞ ⊖ Expected Patterns

⊟ ⊕ Unexpected Patterns

- Need to combine expectation of users with evidence from data (i.e., extracted patterns)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

There is an alternate criteria known as unexpectedness which says that if you already have some knowledge and some evidence, what is the value of the new knowledge, what is the novelty of the new knowledge new rule. So, these you can this Venn diagram illustrates that you can have this kind of situations, which are for the extracted patterns. So, I stop my discussion on association rule here.

So, in summary what you have learnt is to find out define what is an association rule, how to apply support confidence measures, how to apply the apriori algorithm to get frequent item sets, and then how to apply the rule generation using the apriori on the frequent item sets to get rules. Then we have studied different interestingness measures like Jaccard coefficient, like PS values, like lift to evaluate the extracted rules which you can apply. In my later part in the course, when we study visualization will study some of these visualization techniques for rules. And finally, we defined that it is not just interesting and valid it also has to be new novel with respect to the current domain. So, this closes my discussion on association rules.

In my next slide next lectures, I will go into another class of models which are predictive like it helps you classify. So, given a new situation, for example, I am now telling people who buy this also buy another item. Now, I can see if a new customer comes without compromise profile of the customer, can I predict what he will buy, can I predict what he will buy based on certain previous observations, so that forms my next topic on classifications, classification rules.

So, thank you for today I will cover the next in the next chapter.