

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Pattern Evaluation: Which Patterns Are Interesting?

Lecture 15 · Friday 16 October 2026

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–4 min) Recap: Apriori and FP-Growth find frequent itemsets
2. (4–14 min) When strong rules mislead: the tea and coffee example
3. (14–26 min) Lift and χ^2, computed on our baskets
4. (26–38 min) Null transactions; all-confidence, cosine, Kulczynski, imbalance ratio
5. (38–50 min) Choosing a measure; exercise; summary

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Explain	why high support and confidence can still describe a misleading rule
Compute	lift and χ^2 from a 2×2 contingency table
Define	a null transaction and a null-invariant measure
Compute	all-confidence, cosine, Kulczynski and the imbalance ratio
Choose	an appropriate measure for a given dataset

Pattern Evaluation: Which Patterns Are Interesting?

IIT Madras Zanzibar

→ **Why Support and Confidence Are Not Enough**

Lift and χ^2

Null Transactions and Null-Invariant Measures

Choosing a Measure

In-Class Exercise

When a Strong Rule Misleads: Tea and Coffee

IIT Madras Zanzibar

	coffee	no coffee	total
chai	150	50	200
no chai	750	50	800
total	900	100	1,000

1. **support(chai → coffee)**
   150 / 1,000 = 15%

2. **confidence**
   150 / 200 = 75% — looks strong

3. **but P(coffee) = 90%**
   lift = 0.75 / 0.90 = 0.83 < 1

The trap: chai drinkers are LESS likely than average to drink coffee. Confidence alone hides this.

On our data: the same trap is rice seed → hoe: confidence 55%, but hoe is in 60% of baskets (lift 0.91).

Pattern Evaluation: Which Patterns Are Interesting?

IIT Madras Zanzibar

Why Support and Confidence Are Not Enough

→ **Lift and χ^2**

Null Transactions and Null-Invariant Measures

Choosing a Measure

In-Class Exercise

The Contingency Table: The Four Counts Behind Every Rule

IIT Madras Zanzibar

	urea	no urea	total
rice seed	13,043	3,846	16,889
no rice seed	594	**102,517**	103,111
total	13,637	106,363	120,000

1 **both A and B**
13,043 baskets

2 **only one of them**
3,846 and 594

! **NEITHER (null transactions)**
102,517 baskets

Null transaction: a basket containing NEITHER item. They usually vastly outnumber the rest — and some measures are swayed by them.

Lift: Observed Together vs Expected by Chance

$$\text{lift}(A, B) = P(A \cup B) \,/\, (\, P(A) \times P(B)\,)$$

observed together ÷ expected if independent

pair	P(A)	P(B)	expected P(A)P(B)	observed P(A ∪ B)	lift
rice seed, urea	0.141	0.114	0.0160	0.1087	**6.80**
clove seedlings, shade net	0.201	0.162	0.0326	0.1557	**4.77**
sprayer, pesticide	0.099	0.143	0.0142	0.0629	**4.43**
rice seed, hoe	0.141	0.600	0.0844	0.0771	**0.91**

Read: lift 6.80 = rice-seed buyers buy urea 6.8 times as often as chance predicts. 0.91 = slightly LESS than chance.

χ^2 (Chi-Squared): Is the Association Statistically Real?

IIT Madras Zanzibar

$$\chi^2 = \Sigma\ (\text{observed} - \text{expected})^2\ /\ \text{expected}$$

over the 4 cells; > 3.84 ⇒ significant at 5% (1 degree of freedom)

pair	χ^2	lift	direction
rice seed, urea	84,649	6.80	positive
rice seed, hoe	224	0.91	**negative**

1 **χ^2 answers**
'are A and B independent?' — not 'together or apart?'

2 **Both are significant**
84,649 and 224 ≫ 3.84

! **Check the direction**
use lift: hoe is significantly LESS likely

Big-data caution: with 120,000 baskets almost any difference is 'significant'. Significant ≠ large ≠ useful.

Pattern Evaluation: Which Patterns Are Interesting?

IIT Madras Zanzibar

Why Support and Confidence Are Not Enough

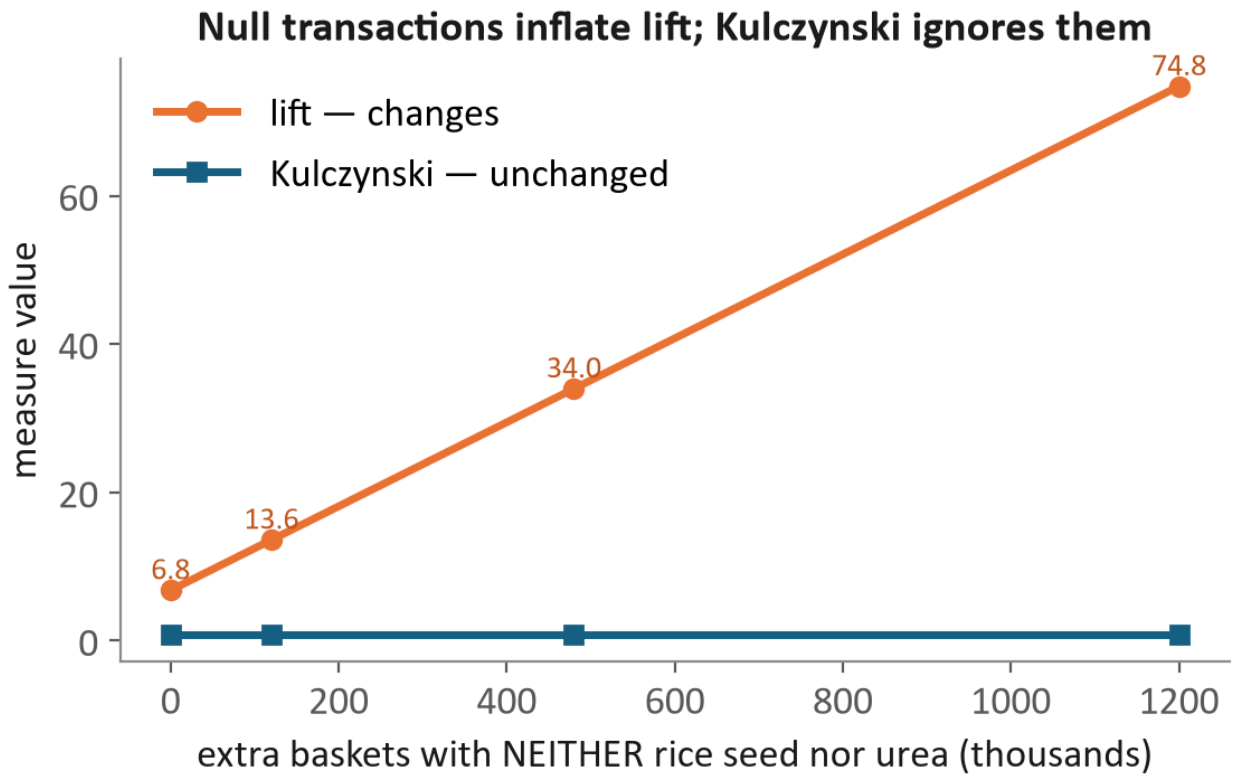
Lift and χ^2

→ **Null Transactions and Null-Invariant Measures**

Choosing a Measure

In-Class Exercise

The Null-Transaction Problem



extra null baskets	lift	Kulczynski
0	6.8	0.864
120,000	13.6	0.864
480,000	34.0	0.864
1,200,000	74.8	0.864

Null-invariant: a measure whose value does not change when baskets containing neither item are added.

Why it matters: real stores sell thousands of products, so for any one pair almost every basket is a null transaction.

Four Null-Invariant Measures

IIT Madras Zanzibar

Measure	Formula	rice seed, urea	rice seed, hoe
all-confidence	min(P(B|A), P(A|B))	0.772	0.128
max-confidence	max(P(B|A), P(A|B))	0.956	0.548
Kulczynski	½ (P(B|A) + P(A|B))	**0.864**	**0.338**
cosine	count(A ∪ B) / √(count(A) × count(B))	0.859	0.265

1 **P(B|A) and P(A|B)**
rice seed & urea: 13,043/16,889 = 0.772 and 13,043/13,637 = 0.956

2 **Kulczynski**
½ (0.772 + 0.956) = 0.864 → strong

All four: range 0 to 1 and ignore the 'neither' cell — so they are null-invariant.

Imbalance Ratio: Is One Item Much More Popular?

$$\mathrm{IR}(A, B) = |\,\mathrm{count}(A) - \mathrm{count}(B)\,| \,/\, (\,\mathrm{count}(A) + \mathrm{count}(B) - \mathrm{count}(A \cup B)\,)$$

0 = balanced popularity; near 1 = one item much more popular

pair	count(A)	count(B)	Kulczynski	IR	reading
rice seed, urea	16,889	13,637	0.864	0.19	balanced, strongly together
sprayer, pesticide	11,889	17,208	0.537	0.25	balanced, moderately together
rice seed, hoe	16,889	72,003	0.338	0.69	very unbalanced; weak

Best practice: report Kulczynski together with the imbalance ratio — the pair of numbers tells you both strength and balance.

Pattern Evaluation: Which Patterns Are Interesting?

IIT Madras Zanzibar

Why Support and Confidence Are Not Enough

Lift and χ^2

Null Transactions and Null-Invariant Measures

→ **Choosing a Measure**

In-Class Exercise

Choosing a Measure

IIT Madras Zanzibar

Measure	Null-invariant?	Range	Use it for
support, confidence	—	0 – 1	finding candidates (thresholds)
lift	no	0 – ∞, 1 = independent	quick check: above or below chance
χ^2	no	0 – ∞	is the association statistically real?
Kulczynski + IR	yes	0 – 1	large sparse data (stores with many products)
all-confidence / cosine	yes	0 – 1	conservative strength check

Our recommendation: use support/confidence to find rules; lift to check direction; Kulczynski + IR to rank them on big sparse data.

Pattern Evaluation in the Real World

IIT Madras Zanzibar

1 Retail

rank thousands of rules; act only on lifted, balanced ones

2 Recommendations

avoid recommending what everyone buys anyway (the hoe effect)

3 Healthcare

drug pairs: significance (χ^2) AND direction (lift)

4 Fraud detection

rare patterns with high Kulczynski can flag collusion

5 Marketing

measure if a promotion truly lifts a second product

6 Agro-dealers (our data)

bundle NPK + cassava cuttings; ignore hoe rules

Common thread: a frequent pattern is only useful if it is more than popularity — measure it properly.

Pattern Evaluation: Which Patterns Are Interesting?

IIT Madras Zanzibar

Why Support and Confidence Are Not Enough

Lift and χ^2

Null Transactions and Null-Invariant Measures

Choosing a Measure

→ **In-Class Exercise**

In-Class Exercise (8 minutes, individually)

IIT Madras Zanzibar

IN-CLASS TASK

10,000 baskets: milk in 6,000; bread in 1,000; both in 800.

1. Build the 2 × 2 contingency table (both / milk only / bread only / neither).
2. Compute confidence(bread → milk) and lift.
3. Compute Kulczynski and the imbalance ratio.
4. Add 90,000 baskets with neither item. Which of lift and Kulczynski changes?

Solution

IIT Madras Zanzibar

1 **Table**
both 800 · milk only 5,200 · bread only 200 · neither 3,800

2 **Confidence, lift**
800/1,000 = 0.80; lift = 0.08 / (0.6 × 0.1) = 1.33

3 **Kulczynski, IR**
½(800/1,000 + 800/6,000) = ½(0.80 + 0.133) = 0.467; IR = 5,000 / 6,200 = 0.81 (unbalanced)

4 **Add 90,000 nulls**
n = 100,000: lift = 0.008 / (0.06 × 0.01) = 13.3 (changed ×10); Kulczynski 0.467 (unchanged)

Summary & Exam Checkpoints

IIT Madras Zanzibar

Confidence can mislead

Chai → coffee: 75% confidence but lift 0.83. Popular items create false rules.

Lift and χ^2

Lift: observed / expected (>1 together). χ^2: real or not, but not direction. Both swayed by null transactions.

Null-invariant

all-confidence, max-confidence, Kulczynski, cosine: ignore baskets with neither item.

Report Kulc + IR

Rice seed & urea: Kulc 0.86, IR 0.19. Rice seed & hoe: Kulc 0.34, IR 0.69.

What's Next

Wk 6 **Classification begins**
decision trees (Wed), Naive Bayes (Thu + Tutorial 6), overfitting and scaling trees (Fri)

Today **Mini-project**
dataset approval checkpoint — submit by 11:59 pm

References & Further Reading

- Tan, P.-N., Kumar, V. & Srivastava, J. (2004). Selecting the right objective measure for association analysis. Information Systems, 29(4), 293–313.
- Wu, T., Chen, Y. & Han, J. (2010). Re-examination of interestingness measures in pattern mining: a unified framework. Data Mining and Knowledge Discovery, 21, 371–397.
- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed. — Chapter 4.4 (pattern evaluation).