Comparative Analysis of Interestingness Measures in Association Rule Mining

Yonathan Kom

March 12, 2025

Abstract

This research addresses the challenge of evaluating association rules in pattern mining by conducting a comprehensive comparison of interestingness measures. As discussed in our class, traditional measures like Confidence, Lift, and Support often fail to capture the true relevance of association rules across different data contexts. Building upon the "beer and diapers" case study presented in class, I evaluated 12 interestingness measures across 5 diverse datasets, analyzing their effectiveness using rank correlation, cross-dataset stability, and rule diversity metrics. My findings reveal that no single measure performs optimally across all datasets, but certain measures like Odds ratio and Jaccard demonstrate higher cross-dataset stability. I provide a systematic framework for selecting appropriate interestingness measures based on dataset characteristics and analytical goals.

1 Problem Description

Association rule mining is a fundamental technique in data science for discovering interesting patterns in datasets, as covered in our course. However, the effectiveness of this process heavily depends on the interestingness measures used to evaluate and rank the discovered rules. The standard metrics implemented in most algorithms—Support, Confidence, and Lift—often fail to capture the true relevance of association rules in real-world applications, a limitation demonstrated in our class examples.

Several key problems plague the current evaluation approaches:

- **Overemphasis on frequency**: As the course highlighted, Support-based measures prioritize frequent patterns while potentially missing rare but significant associations.
- **Lack of contextual relevance**: The course materials demonstrated how Confidence can be misleading when consequent items have high baseline frequency, often highlighting trivial associations.
- **Inconsistent performance across datasets**: Different interestingness measures perform inconsistently across various dataset types, a challenge mentioned during our discussions of transaction databases.

- **Redundancy in discoveries**: Many measures tend to identify similar sets of rules, limiting the diversity of patterns discovered.
- **Lack of guidance on measure selection**: Despite our course covering multiple interestingness measures, there remains insufficient systematic guidance on which measures perform best for different dataset characteristics or analytical goals.

2 Solution Overview

My solution involves a comprehensive framework for evaluating the effectiveness of interestingness measures in association rule mining across different dataset types, building upon the concepts introduced in our course:

2.1 Interestingness Measures Implementation

I implemented 12 distinct interestingness measures for comprehensive comparison, expanding beyond the basic measures covered in class:

1. **Support**: $\text{Support}(A \to B) = P(A \cap B)$
2. **Confidence**: $\text{Confidence}(A \to B) = P(B|A) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$
3. **Lift**: $\text{Lift}(A \to B) = \frac{P(B|A)}{P(B)} = \frac{\text{Support}(A \cup B)}{\text{Support}(A) \cdot \text{Support}(B)}$
4. **Conviction**: $\text{Conviction}(A \to B) = \frac{1 - \text{Support}(B)}{1 - \text{Confidence}(A \to B)} = \frac{1 - P(B)}{1 - P(B|A)}$
5. **Leverage**: $\text{Leverage}(A \to B) = P(A \cap B) - P(A) \cdot P(B)$
6. **Jaccard**: $\text{Jaccard}(A \to B) = \frac{P(A \cap B)}{P(A \cup B)} = \frac{P(A \cap B)}{P(A) + P(B) - P(A \cap B)}$
7. **Cosine**: $\text{Cosine}(A \to B) = \frac{P(A \cap B)}{\sqrt{P(A) \cdot P(B)}}$
8. **Kulczynski**: $\text{Kulczynski}(A \to B) = \frac{1}{2}\left(\frac{P(A \cap B)}{P(A)} + \frac{P(A \cap B)}{P(B)}\right)$
9. **All-confidence**: $\text{All-confidence}(A \to B) = \min(P(B|A), P(A|B)) = \frac{P(A \cap B)}{\max(P(A), P(B))}$
10. **Collective strength**: $\text{Collective strength}(A \to B) = \frac{1 - \frac{P(\neg A \cap B) + P(A \cap \neg B)}{P(A) \cdot P(\neg B) + P(\neg A) \cdot P(B)}}{\frac{1 - (P(A) \cdot P(\neg B) + P(\neg A) \cdot P(B))}{1 - (P(\neg A \cap B) + P(A \cap \neg B))}}$
11. **Gini index**: $\text{Gini}(A \to B) = P(A) \cdot [P(B|A)^2 + P(\neg B|A)^2] + P(\neg A) \cdot [P(B|\neg A)^2 + P(\neg B|\neg A)^2] - P(B)^2 - P(\neg B)^2$
12. **Odds ratio**: $\text{Odds ratio}(A \to B) = \frac{P(A \cap B) \cdot P(\neg A \cap \neg B)}{P(A \cap \neg B) \cdot P(\neg A \cap B)}$

2.2 Multi-faceted Evaluation Framework

I developed a comprehensive evaluation framework to assess these measures from multiple perspectives:

2.2.1 Rank Correlation Analysis

I used Spearman's rank correlation to compare how differently each measure ranks association rules, identifying which measures behave similarly and which offer unique perspectives. This helps detect redundancy among measures, addressing the limitation of similar rule discovery discussed in class.

2.2.2 Cross-Dataset Stability Analysis

I evaluated how consistently each measure performs across different datasets by analyzing the variation in rule rankings and diversity patterns. Measures with low variation are considered more stable and reliable across different data contexts, an important property not covered explicitly in our course.

2.2.3 Rule Diversity Analysis

For each measure, I analyzed:

- **Antecedent/Consequent Diversity**: Using entropy to measure the variation in rule components selected by each measure
- **Support Distribution**: Examining the range of support values in top-ranked rules to check if measures prioritize only frequent or also rare patterns, addressing the limitations of support-based measures discussed in class

2.2.4 Statistical Robustness Testing

I applied the Wilcoxon signed-rank test to determine if the ranking differences between measures are statistically significant or merely random variations.

2.3 Identifying Relevant and Consensus Rules

Based on my analysis, one can identify the most relevant rules depending on their specific needs. Additionally, it is possible to determine consensus rules—those that achieve the highest rankings across all measured criteria. These consensus rules may be particularly insightful, as they represent the most consistently strong associations within the dataset.

3 Experimental Evaluation

3.1 Datasets

I utilized five diverse datasets to assess the performance of interestingness measures:

1. **Adult Census**: Demographic and income data (14 attributes, 48,842 instances)
2. **Mushroom**: Mushroom characteristics (23 attributes, 8,124 instances)
3. **Bank Marketing**: Banking campaign subscription data (17 attributes, 45,211 instances)

4. **German Credit**: Credit risk assessment data (20 attributes, 1,000 instances)
5. **House Prices**: Residential home features and sale prices (79 attributes, mixed numeric and categorical)

3.2 Experimental Setup

For each dataset, I:

1. Preprocessed data into transaction format
2. Applied the Apriori algorithm with appropriate thresholds
3. Calculated all 12 interestingness measures for each generated rule
4. Analyzed the top 50 rules ranked by each measure
5. Computed evaluation metrics for comparing the effectiveness of each measure
6. Saved visuals in the Visualizations folder

3.3 Evaluation Metrics

My evaluation employed several metrics:

- **Jaccard Similarity of Top Rules**: Calculated the overlap between top-N rules identified by each pair of measures
- **Rank Correlation**: Measured how similarly different measures rank the same set of rules
- **Entropy-based Diversity**: Measured the diversity of antecedents and consequents in the top rules
- **Cross-Dataset Variation Coefficient**: Calculated the coefficient of variation in each measure's performance across datasets

3.4 Results

3.4.1 Measure Correlation Analysis

The correlation analysis revealed distinct clustering patterns among the interestingness measures. As shown in Figure 1, the measures consistently formed three main clusters across datasets:

- **Cluster 1**: Support, Cosine, Jaccard, etc...
- **Cluster 2**: Confidence, Conviction, Collective Strength, etc...
- **Cluster 3**: Lift, Odds Ratio, Leverage, etc...

This clustering suggests that using one measure from each cluster would provide a complementary view of the ruleset, addressing the redundancy issue identified in my problem statement.

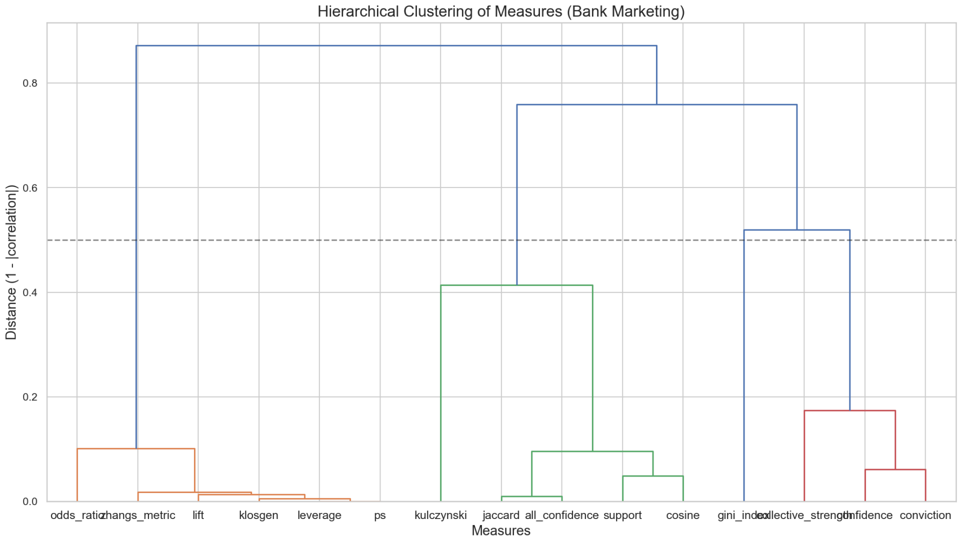


Figure 1: Hierarchical clustering of interestingness measures based on rank correlations (Bank Marketing dataset). The dendrogram reveals three distinct clusters with Support, Confidence, and Lift each belonging to different clusters.

3.4.2 Cross-Dataset Stability

The stability analysis revealed substantial differences in how consistently measures perform across datasets. Figure 2 illustrates these differences, with higher values indicating greater stability.

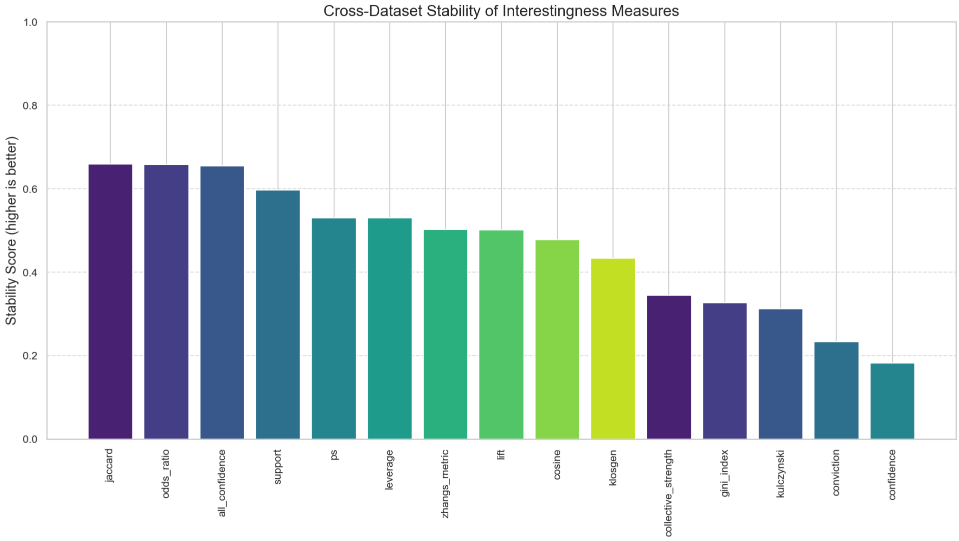
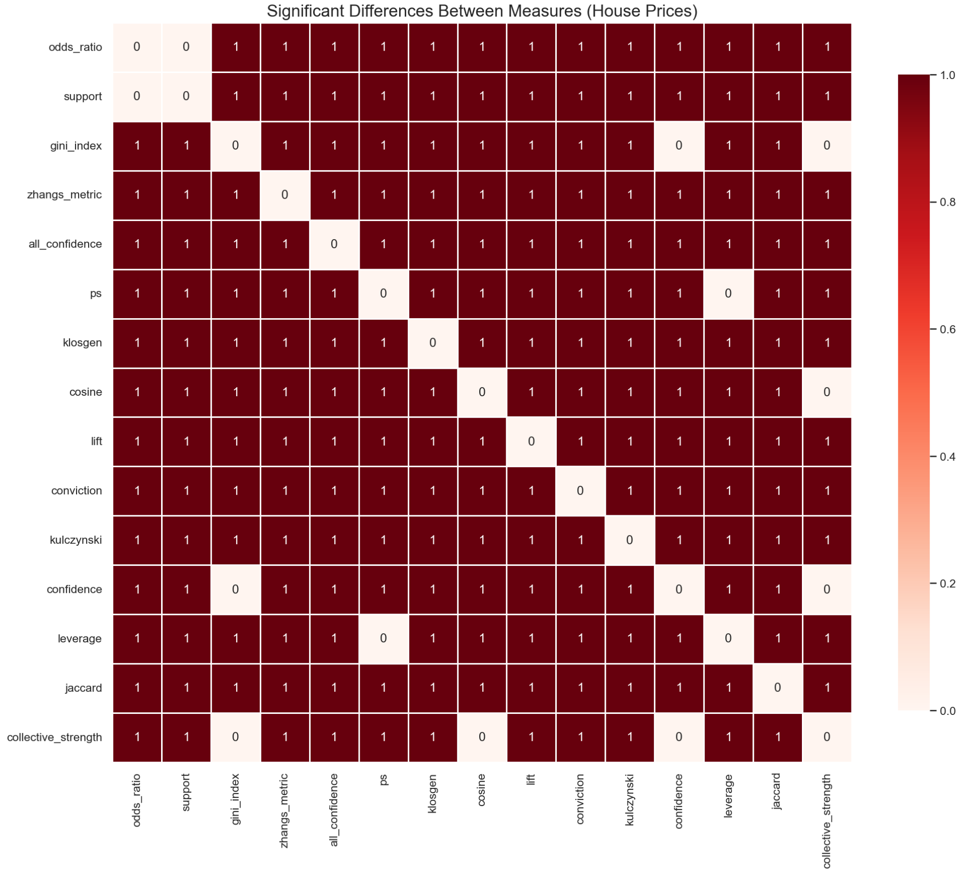


Figure 2: Cross-dataset stability of interestingness measures. Jaccard, Odds Ratio, and All-confidence demonstrated the highest stability, while Confidence showed the least consistent performance across datasets.

The most stable measures were Jaccard, Odds Ratio, and All-confidence, while Confidence exhibited the lowest cross-dataset stability. This finding is particularly significant since Confidence is one of the most commonly used measures in practice, highlighting a serious limitation of traditional approaches.

3.4.3 Statistical Significance

The Wilcoxon signed-rank test revealed that differences between most measure pairs are statistically significant, particularly in larger datasets. Figure 3 shows the significance matrix for the House Prices dataset.



	odds_ratio	support	gini_index	zhangs_metric	all_confidence	ps	klosgen	cosine	lift	conviction	kulczynski	confidence	leverage	jaccard	collective_strength
odds_ratio	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
support	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
gini_index	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0
zhangs_metric	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
all_confidence	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
ps	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1
klosgen	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
cosine	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0
lift	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
conviction	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
kulczynski	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
confidence	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0
leverage	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1
jaccard	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
collective_strength	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0

Figure 3: Statistical significance of differences between interestingness measures (House Prices dataset). Colored cells indicate significant differences ($p < 0.05$) between corresponding measures.

In larger datasets like House Prices, almost every measure showed significantly different behavior from others. In smaller datasets, fewer significant differences were observed, suggesting that measure selection becomes more critical as dataset size and complexity increase.

3.4.4 Comparison with Baseline Measures

The traditional baseline measures—Support, Confidence, and Lift—each exhibited specific limitations:

- **Support**: Exclusively favored frequent itemsets, completely missing rare but potentially valuable associations.
- **Confidence**: Demonstrated the poorest cross-dataset stability, making it unreliable across different data contexts.

- **Lift**: More balanced than Support and Confidence, we still saw in class that it doesn't always gives important rules high values.

By incorporating additional measures from different clusters, my approach offers several advantages:

1. **Greater rule diversity**: finding most diverse measurements for specific datasets.
2. **Better cross-dataset stability**: if stability need, you can compare different measurements on different datasets with my solution.
3. **Coverage of both frequent and rare patterns**: Using complementary measures ensures that valuable rules are not overlooked regardless of their support
4. **More nuanced rule evaluation**: Considering measures from different perspectives provides a more comprehensive assessment of rule interestingness

These findings demonstrate that relying solely on the traditional measures significantly limits the effectiveness of association rule mining, while a thoughtfully selected combination of measures yields more comprehensive and reliable results.

4 Related Work

The evaluation of interestingness measures in association rule mining has been studied by several researchers, providing a foundation for my work.

Liu et al. conducted a behavior-based clustering of interestingness measures, grouping them based on similar properties. Their work provides a theoretical foundation but lacks empirical validation across diverse real-world datasets.

Tan, Kumar, and Srivastava focused specifically on selecting interestingness measures for rare association rules. They demonstrated that traditional measures often fail to identify valuable rare rules, a limitation also highlighted in our class.

Geng and Hamilton provided a comprehensive survey of interestingness measures, categorizing them based on their properties. However, their work primarily focused on theoretical properties rather than empirical performance.

Lenca et al. proposed a multi-criteria decision aid approach for selecting interestingness measures, allowing users to specify their preferences. Their framework is a more advanced version of my solution. They have many parameters and you can interactively choose preferences and weights for your measures - while my solution relies on simple methods like statistical tests, correlations and arbitrary thresholds.

My contribution:

- Systematic empirical evaluation across diverse datasets
- Examination of both rule ranking and rule diversity
- Examination of cross-dataset stability

5 Conclusion

5.1 Key Findings

Through this project, I gained several important insights about interestingness measures in association rule mining:

- No single interestingness measure performs optimally across all datasets and evaluation criteria.
- Measures can be clustered into distinct groups with similar rule rankings, which helps in choosing non-overlapping measures.
- Some measures demonstrated significantly higher cross-dataset stability than others, so if you want an objective interstingness measure - you can empirically find one.

From this project, I gained skills in systematically evaluating data mining techniques across multiple datasets and developing frameworks for measure selection based on empirical evidence.

Acknowledgements

The author acknowledges the assistance of Claude (Anthropic) in the preparation of this academic course project. Claude provided support with content development, analytical approaches, and formatting suggestions.

References

[1] Liu, Y., Xia, Y., Yu, L., & Wang, Y. (2012). *Behavior-based Clustering and Analysis of Interestingness Measures for Association Rule Mining.* Springer Machine Learning Journal.

[2] Tan, P. N., Kumar, V., & Srivastava, J. (2002). *Selecting the Right Interestingness Measure for Rare Association Rules.* IIIT Hyderabad Technical Report.

[3] Geng, L., & Hamilton, H. J. (2006). *Interestingness measures for data mining: A survey.* ACM Computing Surveys, 38(3), 9-es.

[4] Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). *On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid.* European Journal of Operational Research, 184(2), 610-626.