Sparsification Levels and Robustness in Contrastive Learning-Based Recommenders

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: To what extent does the sparsification level (e.g., 50% vs. 90% edge reduction) affect the robustness of contrastive learning-based recommenders against adversarial attacks, as measured by metrics. Machine-learning models have demonstrated great success in learning complex patterns that enable them to make predictions about unobserved data. In addition to using models for prediction, the ability to interpret what a model has learned is receiving an increasing amount of. 9 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Definitions, methods, and applications in interpretable machine learning. Research question: To what extent does the sparsification level (e.g., 50% vs. 90% edge reduction) affect the robustness of contrastive learning-based recommenders against adversarial attacks, as measured by metrics like AUC or Hit Ratio?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.3/10.

3 Results

14 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 8.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Machine-learning models have demonstrated great success in learning complex patterns that enable them to make prediction	✓	0.31
The ability to interpret what a model has learned is receiving an increasing amount of attention.	✓	0.22
There is considerable confusion about the notion of interpretability.	✓	0.17
It is unclear how the wide array of proposed interpretation methods are related and what common concepts can be used to	✓	0.30
The PDR framework provides 3 overarching desiderata for evaluation: predictive accuracy, descriptive accuracy, and relev	✓	0.40
We introduce a categorization of existing techniques into model-based and post hoc categories, with subgroups including	✓	0.31
We provide numerous real-world examples to demonstrate how practitioners can use the PDR framework to evaluate and under	✓	0.33
These examples highlight the often underappreciated role played by human audiences in discussions of interpretability.	✓	0.27
Based on our framework, we discuss limitations of existing methods and directions for future research.	✓	0.21

References

- https://doi.org/10.1214/21-ss133
- https://doi.org/10.1109/access.2020.3041951
- https://doi.org/10.1073/pnas.1900654116