Technological University Dublin
ARROW@TU Dublin

SAML-25 Workshop on Statistical and Machine Learning

Research Institutes/Centres/Groups

2025-06-05

Improving Node Classification for Graphs withWeak Feature Signals: A Similarity-Entropy Aggregation Approach

Brian Daniel Bernhardt
University of Cassino and Southern Lazio, briandaniel.bernhardt@unicas.it

Chiara Marciano
University of Cassino and Southern Lazio, chiara.marciano@unicas.it

Mario Rosario Guarracino
University of Cassino and Southern Lazio, mariorosario.guarracino@unicas.it

Follow this and additional works at: https://arrow.tudublin.ie/saml

 Part of the Statistics and Probability Commons

Recommended Citation

Bernhardt, Brian Daniel; Marciano, Chiara; and Guarracino, Mario Rosario, "Improving Node Classification for Graphs withWeak Feature Signals: A Similarity-Entropy Aggregation Approach" (2025). *SAML-25 Workshop on Statistical and Machine Learning*. 13.
https://arrow.tudublin.ie/saml/13

This Article is brought to you by the EUt+ Academic Press a free to read and publish press of the European University of Technology.



This work is licensed under a Creative Commons Attribution-NonCommercial-Share Alike 4.0 International License.

Improving Node Classification for Graphs with Weak Feature Signals: A Similarity-Entropy Aggregation Approach

Brian Daniel Bernhardt
University of Cassino and Southern Lazio
Cassino, Italy
Indatis s.r.l.
Cassino, Italy
briandaniel.bernhardt@unicas.it

Chiara Marciano
University of Cassino and Southern Lazio
Cassino, Italy
chiara.marciano@unicas.it

Mario Rosario Guarracino
University of Cassino and Southern Lazio
Cassino, Italy
Laboratory of Algorithms and Technologies for Networks Analysis, Higher School of Economics
Nizhny Novgorod, Russia
mariorosario.guarracino@unicas.it

Abstract

Graph Neural Networks (GNNs) have established themselves as powerful tools for graph-structured data. However, when feature separability among nodes is low, conventional neighborhood aggregation strategies often result in performance degradation due to over-smoothing and noisy information propagation. In this work, we introduce a novel GNN framework that refines the aggregation process by integrating feature similarity and neighborhood entropy into node message passing. Unlike standard models that uniformly aggregate neighbor information, this new model dynamically adjusts neighbor influence, prioritizing nodes with high similarity and low entropy. We evaluate the model on synthetic graphs generated using the Stochastic Block Model (SBM), varying homophily and class imbalance conditions to simulate challenging classification scenarios. Our experiments demonstrate that our method consistently outperforms the GraphSAGE model across Accuracy, Balanced Accuracy, F1 Score, and Matthews Correlation Coefficient, particularly in graphs with weak class signal and high structural noise. Statistical significance tests confirm the robustness of the observed improvements. Furthermore, the optimization of aggregation weights provides information about the model's adaptation to different graph structures, enhancing interpretability. These findings suggest that integrating structural information selectively during aggregation can significantly improve GNN performance in complex graph environments, offering a new course toward more explainable graph-based models.

Keywords

Graph Neural Networks, Supervised Learning, Node Aggregation

1 Challenges in Graph Neural Networks and Our Approach

Graph Neural Networks (GNNs) have become a reliable tool in analyzing graph-structured data, successfully applying message-passing mechanisms where nodes iteratively aggregate features from their neighbors [4]. However, this strategy presents challenges when node features lack strong discriminative power. Phenomena such as over-smoothing [1] and the propagation of noisy neighborhood information can weaken classification performance, particularly in graphs where feature separability is low. Traditional aggregation methods, often based on uniform or mean pooling of neighbor features, are insufficient in these complex scenarios.

Recent studies have highlighted that in graphs where neighboring nodes differ significantly, conventional aggregation can introduce detrimental noise rather than useful information [5]. Moreover, enforcing aggregation based solely on network connectivity may lead GNNs to perform worse than simple multilayer perceptrons, especially as depth increases. To address these issues, we propose a new approach, which introduces an adaptive aggregation mechanism grounded on two key concepts: feature similarity and neighborhood entropy. The proposed model operates by assigning dynamic weights to each neighboring node during aggregation. Nodes that exhibit high similarity in features and low neighborhood entropy are given greater influence in updating the target node's representation. Similarity is measured through the mean squared difference between feature vectors, while entropy captures the heterogeneity of a node's local environment, based on Shannon entropy. By minimizing the impact of uncertain or structurally noisy nodes, our approach aims to refine the information passed during message propagation.

The aggregation mechanism is implemented within a two-layer GraphSAGE framework [2]. The first layer employs a standard mean aggregator to stabilize initial feature representations, while the second layer incorporates the custom similarity-entropy weighting scheme. This combination ensures both generalization and fine-grained adaptation to local graph structures. Furthermore, the relative importance of similarity and entropy is governed by tunable parameters, optimized using Bayesian optimization to adapt to different structural properties of the graph. This design not only enhances performance but also increases model interpretability by explicitly exposing how aggregation decisions are made based on neighborhood characteristics.

2 Experimental Validation and Results

To assess the effectiveness of our approach, we conducted experiments using synthetic datasets generated via the Stochastic Block Model (SBM). The graphs contained 2000 nodes divided into two classes under varying conditions of homophily (high, low) and class balance (balanced, imbalanced). Nodes were assigned five features: four drawn from class-specific Gaussian distributions and one from a uniform distribution to introduce noise, simulating scenarios with low feature separability. In the SBM configuration, high homophily

graphs had strong intra-class connections and weak inter-class links, while low homophily graphs had random connections without preference for class. This setup allowed us to systematically evaluate performance across structured and noisy environments. Four graph variants were generated: high homophily–balanced, high homophily–imbalanced, low homophily–balanced, and low homophily–imbalanced.

Our approach was implemented with two convolutional layers: the first with mean aggregation to stabilize features and the second employing the custom similarity-entropy weighted aggregation. For each graph type, α and β coefficients, controlling the influence of similarity and entropy, respectively, were optimized using Bayesian optimization [3]. Model performance was evaluated over 20 random seeds, with a 50/25/25 train/validation/test split. Metrics included Accuracy, Balanced Accuracy, Weighted F1 Score, and Matthews Correlation Coefficient (MCC), and paired t-tests were conducted to determine statistical significance compared to the baseline GraphSAGE model.

The results consistently showed superior performance of our approach over GraphSAGE across all metrics. In graphs with high homophily and balanced classes, improvements were modest but consistent. More significant gains were observed under challenging conditions: for low homophily–balanced graphs, Balanced Accuracy improved from 90.01% to 92.52%, and for low homophily–imbalanced graphs, from 71.90% to 81.62%. Similarly, in high homophily-imbalanced graphs, our method improved Balanced Accuracy from 91.79% to 95.61%. All improvements were statistically significant, with p-values below 0.01 or 0.001 depending on the setting. Analysis of the optimized coefficients further confirmed the model's adaptive behavior. In high homophily–balanced settings, the model assigned a higher weight to entropy, recognizing the reliability of local neighborhoods. Conversely, in low homophily–imbalanced graphs, similarity became the dominant factor, reflecting the need to rely on structural rather than class-based information.

Overall, the experiments demonstrated that introducing similarity and entropy into the aggregation process enables GNNs to better adapt to graph complexity, particularly when traditional aggregation mechanisms struggle. Our approach thus represents a promising direction for improving the performance and explainability of graph neural networks operating under difficult feature conditions.

References

[1] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3438–3445.

[2] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216* (2017).

[3] Yuxiao Hou, Jintang Zhang, Jian Cheng, Kailing Ma, Ruiwen T. B. Ma, Haibin Chen, and Min Choon Yang. 2022. Measuring and improving the use of graph information in graph neural networks. *arXiv preprint arXiv:2206.13170* (2022).

[4] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2017).

[5] Yiquing Xu, Li Sha, Chen Yang, Raymond Chi-Wing Wong, and Jiawei Han. 2020. When Do GNNs Work: Understanding and Improving Neighborhood Aggregation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization, 4399–4406. doi:10.24963/ijcai.2020/608