# Graph Active Learning: utilizing graphs to boost active learning on tabular data

**Saleem Kheralden**
@ Technion
{khsaleem

**Majd Bishara**
@ Technion
majedbishara

**Lana Haj**
@ Technion
lanahaj}@campus.technion.ac.il

## Abstract

Active learning is an effective technique for training models on datasets that are challenging to label. This approach leverages uncertainty metrics to identify data points that are likely to contribute most to the model's improvement. Among these metrics, entropy is one of the most widely used. While active learning can be both efficient and impactful, traditional methods often select points based solely on the model's predictions, without considering the underlying data distribution. To address this limitation, we propose an enhanced active learning method that incorporates data distribution by using Graph Neural Networks (GNNs) to represent data in graph form, leveraging both uncertainty and structural relationships for a more informed selection process.

## 1 Introduction

Active Learning (AL) is an effective approach for training models on datasets that are difficult to label, using uncertainty metrics to identify and select data points that are most likely to improve the model's performance. Although AL has proven effective, especially in yielding strong results across various tasks, there is potential for enhancing its performance on tabular data. We aim to improve AL by refining the selection process and adding complexity through graph-based methods. Specifically, we will convert tabular data into graph structures and leverage centrality metrics to inform the selection of data points. Building on this, we will incorporate Graph Neural Networks (GNNs) to classify labeled nodes.

Graph Neural Networks (GNNs) have gained significant traction in recent years, achieving remarkable success across tasks such as node, edge, and graph classification. Known for their ability to capture semantic relationships between nodes and their neighbors by aggregating neighboring information, GNNs are particularly well-suited for embedding and classifying graph-structured data.

In this paper, we introduce a novel approach to Active Learning (AL) using GNNs. Our method follows these steps: starting with tabular data, we construct a graph where each data point becomes a node, and edges are defined based on a similarity metric $f$ . We then apply a GNN model to embed the nodes in this graph, training the classification head of the GNN on the embeddings of labeled data points. In parallel, we use centrality metrics $\phi$ to identify a subset of the unlabeled nodes with the highest potential impact if labeled, which are then selected for annotation. With the updated labeled set, we retrain the main classification model and iterate, repeating the graph construction and selection steps with each new batch of labeled data. This approach leverages the structural information in tabular data more effectively, potentially leading to enhanced performance in AL applications.

## 2 Methodology

Let $D = (L, U) \in \mathbb{R}^{n \times m}$ be the data matrix, such that $L$ is the labeled data, $U$ the unlabeled data, Let $f$ be the similarity metric (distance function, euclidean for example), $M$ the classifier (LogisticRegression), and $G$ be GNN model. Let $\mathcal{A}$ be the similarity matrix $\forall i, j \in [n]$, $\mathcal{A}_{i,j} = f(D_i, D_j)$, and $\mathcal{A}^L, \mathcal{A}^U$ be the sub-matrices that corresponds to $L, U$ respectively, let $\delta$ be a hyper-parameter for similarity threshold which is chosen as a quantile from the range of distances calculated between the nodes, $\phi$ is the uncertainty metric where

$$\phi = \alpha \cdot \phi_{entropy} + \beta \cdot \phi_{density} + \gamma \cdot \phi_{centrality}$$

such that $\phi_{entropy}, \phi_{density}, \phi_{centrality}$ are defined similarly to (Hongyun Cai, 2017).

**Algorithm 1** GNN Active Learning

```
function CONSTRUCT_GRAPH ($\mathcal{A}$)
    $E \leftarrow \{(d_i, d_j) : \mathcal{A}_{i,j} < \delta\}$
    return graph $\mathcal{G}(D, E)$
end function

function LABEL_UPDATE ($\mathcal{O}, L, U, U_q$)
    $L_q \leftarrow \mathcal{O}.label(U_q)$
    $L \leftarrow L \cup L_q$
    $U \leftarrow U \setminus L_q$
    return $L, U$
end function

```

**Algorithm 1** GNN Active Learning (Continued)

```
function TRAIN_MODEL
    $M \leftarrow train\ M\ on\ L$
    $G \leftarrow train\ G\ on\ \mathcal{G}.L$
    return $M,\ G$
end function

function RUN_PIPELINE $(D)$
    for $iter \leftarrow \{1, \ldots, iterations\}$ do
        $M \leftarrow$ train_model()
        $\mathcal{G} \leftarrow$ construct_graph($\mathcal{A}$)
        $S \leftarrow \phi(\mathcal{G}.U, U, M, G)$
        $U_q \leftarrow k$ points with the highest $S$
        $L, U \leftarrow label\_update(\mathcal{O}, L, U, U_q)$
    end for
end function
```

## 3 Data

To evaluate our framework, we conducted experiments on five datasets: **Iris** and **Wine Quality**, which are well-established benchmarks in the field, as well as **lab_dataset_2000**, a sample dataset provided during one of the labs. In addition, we created two synthetic datasets, **Clustered** and **Unclustered**, to further test our method. These datasets vary in size, dimensionality, and exhibit differences in a feature that we hypothesize to be important for our framework: *clusteredness*, which we define as the degree to which data points with the same label are spatially close (i.e., form clusters). For the synthetic datasets, we emphasized this characteristic to its extremes, creating both highly clustered and highly dispersed versions of labeled points.

Visualizations of each dataset are provided, with dimensionality reduction using PCA where necessary.

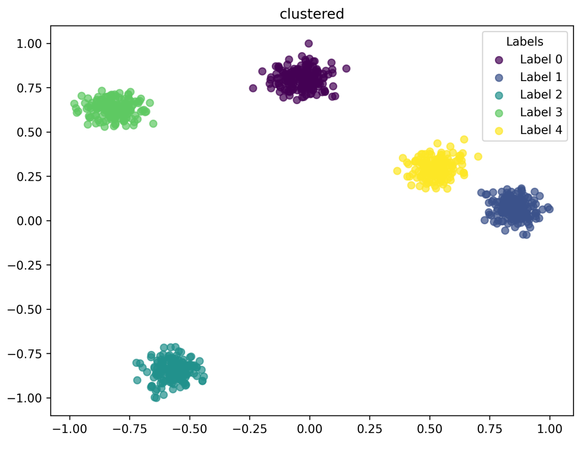

Figure 1: Clustered dataset

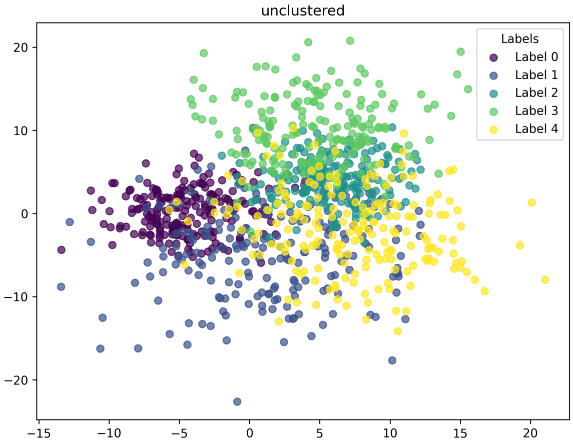

Figure 2: Clustered dataset

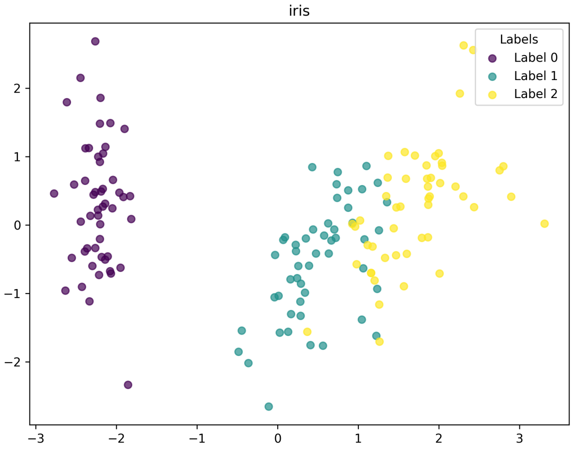

Figure 3: Clustered dataset

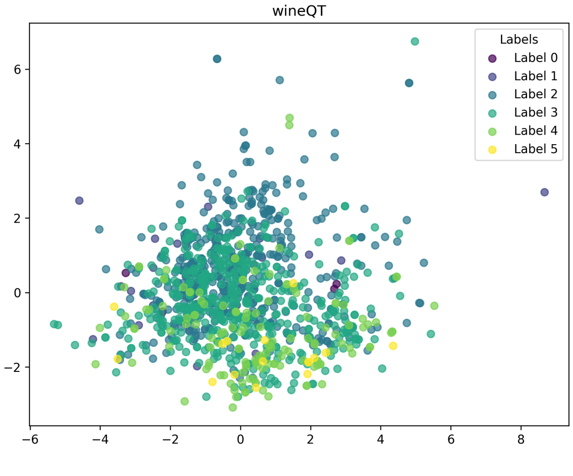

Figure 4: Clustered dataset

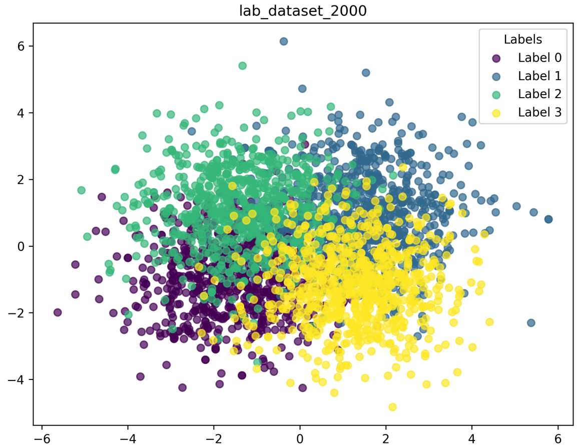

Figure 5: Clustered dataset

## 4 Experimentation and Evaluation

For the experimentation and evaluation, we conducted over 3,000 experiments across the five datasets. For each dataset, we explored a range of hyperparameter combinations, including different graph construction similarity metrics $f$ and threshold values $\delta$, as well as various uncertainty parameters and measures such as PageRank, entropy, and their combinations. Additionally, we experimented with different GNN configurations, including hidden layer sizes and the number of epochs the model was trained for at each iteration. These configurations were applied to our proposed model, **GAL**, and compared with a traditional Active Learning (AL) approach. To evaluate the performance of the models, we computed the accuracy metric (as defined in the course), which served as the primary criterion for comparison.

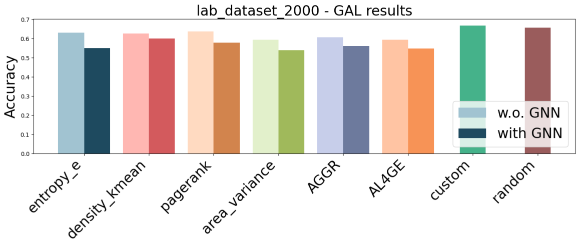

Figure 6: Lab dataset results

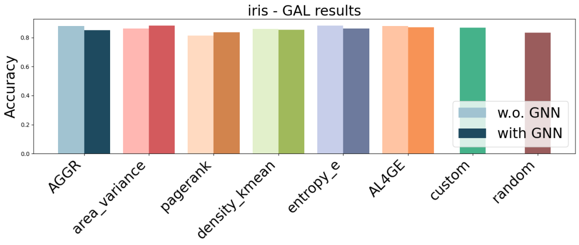

Figure 7: Iris dataset results

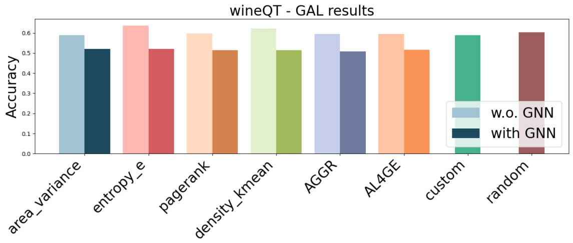

Figure 8: Wine quality dataset results

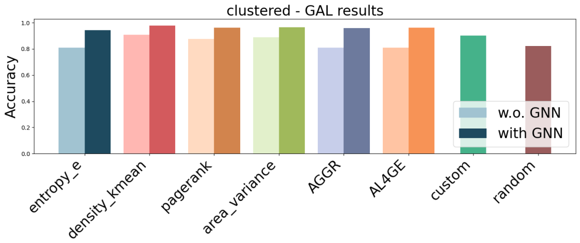

Figure 9: Clustered dataset results

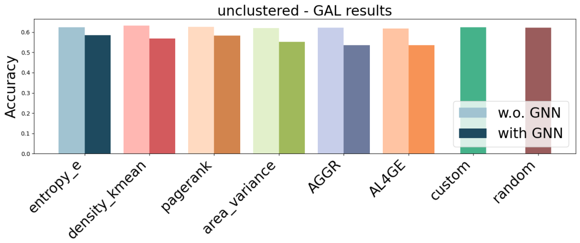

Figure 10: Clustered dataset results

## 5 Related Work

Given our goal of leveraging Graph-based Active Learning (GAL) for tabular data, we examined three methods: Active Learning for Graph Embeddings (AGE) (Hongyun Cai, 2017), Information Gain Propagation (IGP) (Wentao Zhang, 2022), and Active Learning for Graphs with Noisy Structures (GALClean) (Hongliang Chi, 2024). IGP introduces a novel approach by using relaxed queries with soft labels, where the oracle only verifies predicted labels rather than assigning hard labels, thereby maximizing information gain propagation across graph nodes (IGP). (Hongliang Chi, 2024) focuses on noisy graphs, iteratively combining data selection and graph cleaning using a Stochastic Expectation-Maximization framework to counteract structural noise that could otherwise impair model performance (Hongliang Chi, 2024). While both IGP and GALClean provide innovative approaches to improve labeling efficiency, we chose AGE for its ease of implementation and relatively high performance on less complex graph structures. AGE's uncertainty measure has already demonstrated robust results across different domains, making it an ideal choice for our first foray into graph-based AL on tabular data.

# 6 Limitations

Our proposed model, GAL, demonstrated comparable results to the baseline models, indicating its potential for applying graph-based Active Learning (AL) to tabular data. However, several limitations were identified during our initial investigation. Firstly, we applied straightforward and reasonable methods to evaluate the feasibility of the approach, without leveraging more advanced techniques that could enhance performance. Additionally, the GNN component in our model was trained on small datasets, which constrained its ability to reach its full potential. Consequently, the model may not have fully utilized the benefits of graph-based learning, limiting its scalability and performance on larger, more complex datasets.

# 7 Discussion and Conclusions

Despite these limitations, GAL performed similarly to baseline models, highlighting its promise as a tool for Active Learning on tabular data. Notably, the model showed slightly better performance on the clustered datasets compared to the unclustered datasets. This suggests that the inherent structure in clustered data might enhance the performance of graph-based learning methods, which rely on proximity between similar data points. The results encourage further investigation into the applicability of graph-based techniques in Active Learning, particularly in domains where data naturally exhibits clustering behavior.

Looking ahead, there are several directions for improvement. Future work can focus on scaling GAL to larger datasets, where its full potential can be better realized. Furthermore, the incorporation of more sophisticated methods, such as advanced GNN architectures, could improve the model's ability to capture complex relationships in the data. This study lays a solid foundation for GAL's application to a broader range of domains and provides an encouraging direction for future exploration and refinement.

# References

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434.

Yankai Chen Renhe Jiang Weiping Ding Manabu Okumura Dongyuan Li, Zhen Wang. 2024. A survey on deep active learning: Recent advances and new frontiers. *IEEE*.

Suhang Wang Yao Ma Hongliang Chi, Cong Qi. 2024. Active learning for graphs with noisy structures.

Kevin Chen-Chuan Chang Hongyun Cai, Vincent W. Zheng. 2017. Active learning for graph embedding.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2024. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*. Accepted.

Zhenbang You Meng Cao Ping Huang Jiulong Shan Zhi Yang Bin Cui Wentao Zhang, Yexin Wang. 2022. Information gain propagation: A new way to graph active learning with soft labels. *ICLR 2022*.