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, the 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 ϕ 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 $D = (L, U, T) \in \mathbb{R}^{n \times m}$ be the data matrix, such that L is the labeled data, U the unlabeled data and T is the validation 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 $\mathcal{A}^L, \mathcal{A}^U, \mathcal{A}^T$ be the sub-matrices that corresponds to L, U, T respectively, let δ be a hyper-parameter for similarity threshold which is chosen as a quantile from the range of distances calculated between the nodes, ϕ 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)
    $\mathcal{G}^{test} \leftarrow$ construct_graph($\mathcal{A}^T$)
    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
        $L, U \leftarrow label\_update(\mathcal{O}, L, U, U_q)$
        $G.eval(\mathcal{G}^{test})$
        $M.eval(T)$
    end for
end function

All the code is in the GitHub repo BSc-finale

3 Uncertainty Measures

Uncertainty measures are functions applied to the unlabeled data that assign a score to each point, indicating how uncertain the model is about its prediction for that point. These scores guide the selection of points to add to the training dataset in each iteration, prioritizing those that are expected to contribute the most to improving the model's accuracy.
We will use them either seperately or in conjunction with each other.

- **Entropy**: Given data features X, we calculate the output vectors using the classifier and then compute the entropy row-wise on the output. Entropy is calculated as $\sum o_i \log(o_i)$, where o_i represents the model's output probabilities for each class.
- **Density**: We calculate the centroids on X using k-means clustering, with k as a hyperparameter. For each point v_j in X, we find the distance to the closest centroid c_i. The uncertainty is given by $\frac{1}{1+\text{dist}(c_i, v_j)}$, where a smaller distance indicates higher certainty.
- **Area Variance**: For each node v, we create a vector a with dimension equal to the number of labels. Each entry a_i in the vector corresponds to label i and represents the count of v's neighbors that have label i. We then calculate the entropy of vector a to measure the diversity of labels in the node's neighborhood.
- **PageRank**: PageRank is applied to the graph constructed from the dataset to assess the importance of each node (data point) within the graph. Higher PageRank scores suggest points that are more central and influential in the structure.
- **AL4GE**: A composite uncertainty measure that combines *Entropy*, *Density_kmean*, and *PageRank*.
- **AGGR**: An aggregate uncertainty measure that combines *Entropy*, *Density_kmean*, *PageRank*, and *Area Variance*.

4 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 from the 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.

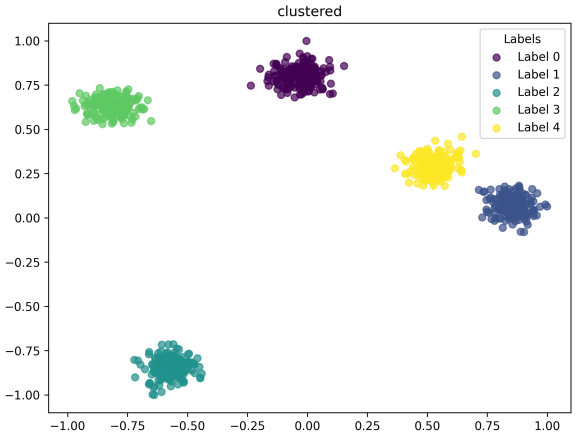

Figure 1: Clustered dataset

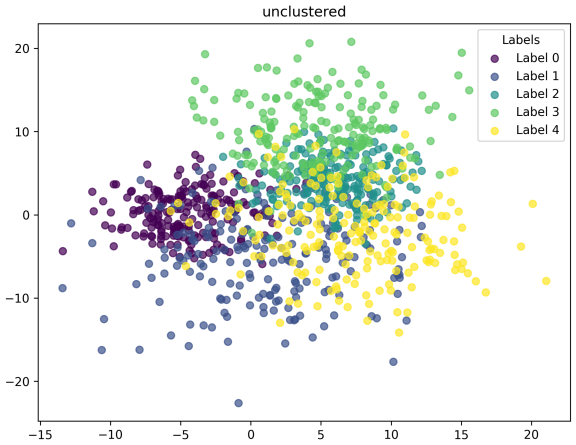

Figure 2: Clustered dataset

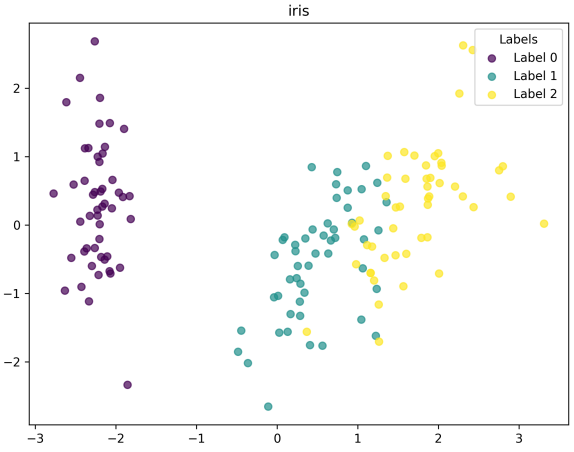

Figure 3: Clustered dataset

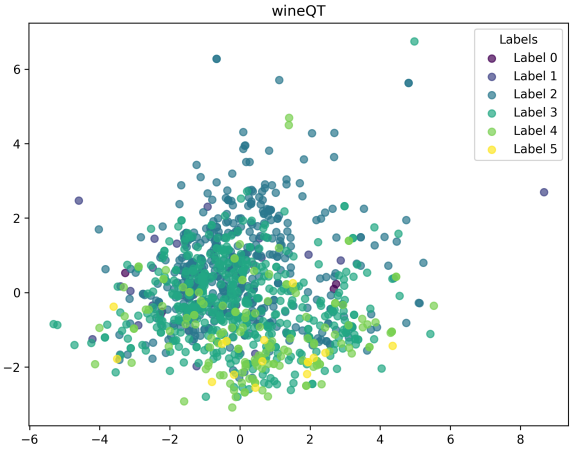

Figure 4: Clustered dataset

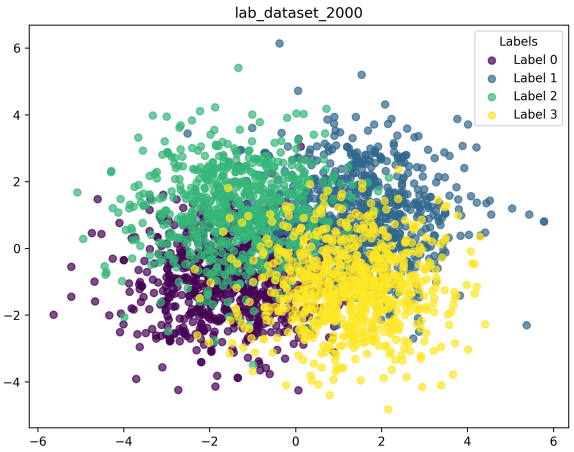

Figure 5: Clustered dataset

5 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 δ, 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 shown in the course), finally we aggregated the runs for a given uncertainty by taking an average of all the configurations

6 Results

The plots below illustrate the performance variations across different uncertainty measures for each dataset. We compare the results of GAL with the GNN and without it. Across most datasets, we observe that the inclusion of the GNN generally performed similarly to or slightly worse than the non-GNN version, except for the "clustered" dataset. As previously noted, this dataset accentuates the clustering characteristic, which we believe is a factor contributing to the superior performance of the GNN in this case.

Regarding the uncertainty measures, results indicate that for the majority of datasets, GAL did not surpass the "custom" or random selection which we consider as a baseline, often achieving similar or lower scores. However, in the case of the clustered dataset, we observed an improvement over the baseline, underscoring the impact of data structure on GAL's performance when using GNNs.

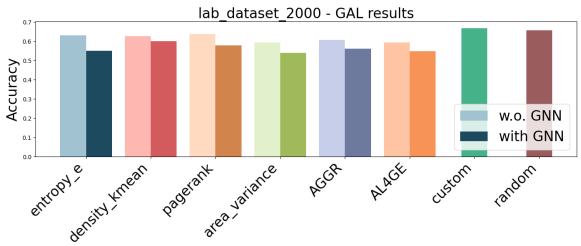

Figure 6: Lab dataset results

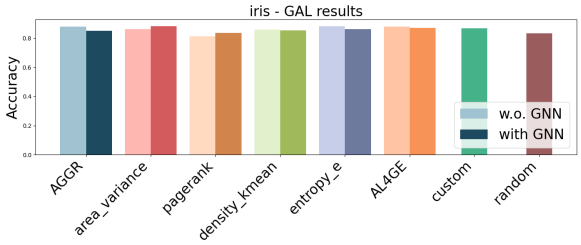

Figure 7: Iris dataset results

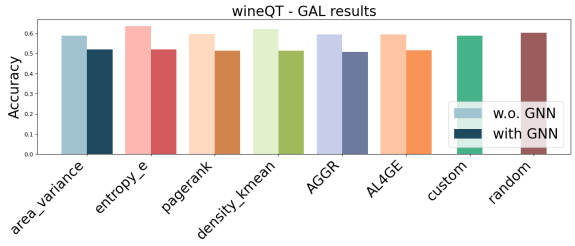

Figure 8: Wine quality dataset results

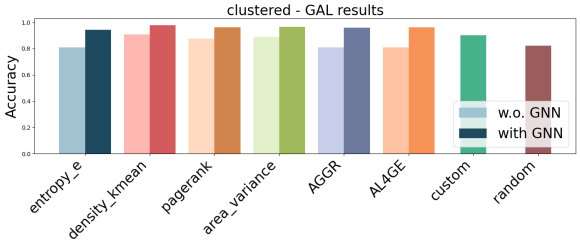

Figure 9: Clustered dataset results

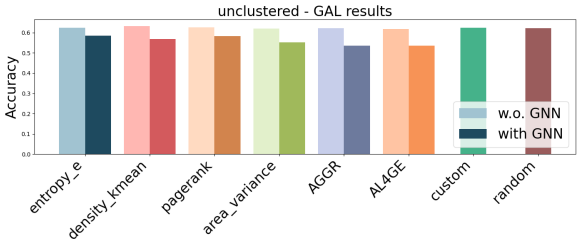

Figure 10: Clustered dataset results

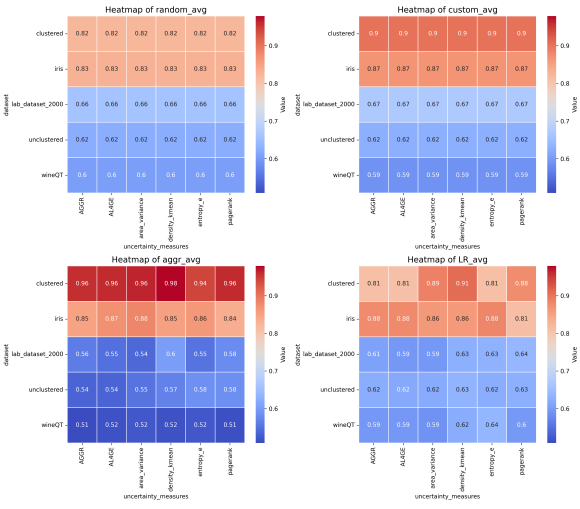

Figure 11: Heatmaps of the results

7 Limitations

Our proposed model, GAL, showed promising results, but several limitations became apparent during this initial investigation. Firstly, our approach was straightforward and did not fully address scaling issues, making graph construction a bottleneck in our experiments. Due to this, we had to limit the dataset size to approximately 3,000 datapoints, constraining the scope of experimentation and limiting the GNN model's potential, as it was trained on relatively small datasets. This size restriction is not ideal for deep learning tasks, particularly with GNNs, which typically benefit from larger datasets. Consequently, the model may not have fully leveraged the advantages of graph-based learning, potentially impacting its scalability and performance on more extensive, complex datasets.

8 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 forward, several directions offer potential for enhancing GAL. Scaling GAL to accommodate larger datasets is a priority, as its full potential may be more evident in such contexts. Strategies such as graph maintenance in a distributed setting or strategic sampling of nodes and edges could mitigate the challenges of building and storing large graphs, enabling GAL to handle more extensive datasets. Additionally, integrating more sophisticated methods, such as advanced GNN architectures, could enhance GAL's ability to capture complex relationships and nuances within the data.

This study establishes a solid foundation for GAL's application across a wider range of domains, highlighting an encouraging path for further exploration and refinement.

9 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.

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*.