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 a method to train a model effectively on a data that's not easy to label, in order to train the model the method utilizes uncertainty metrics with the goal to select points which would have the most contribution to the training process, with entropy uncertainty metric is the most known of those metrics, while this method of training effective and efficient, it selects points based on the output of the trained model without taking into account the distribution of the data, to address this issue we present a new improved AL method, where we utilize the data distribution as well using GNNs and converting the data into graph.

1 Introduction

Active Learning (AL) is a method to train a model effectively on a data that's not easy to label, in order to train the model the method utilizes uncertainty metrics with the goal to select points which would have the most contribution to the training process, while this method proved that it's effective and yielded great results, we think that we can improve upon this method to yield even better performance on tabular data, that is by improving the selection scheme, and adding another layer of complication on this method, we will tackle the selection problem by converting the tabular data into graphs and utilize their centrality metrics, and the complication that we intend to add is GNNs, since we're already gonna work with graphs, we will test different GNN models in order to classify the labeled nodes.

Graph Neural Networks (GNNs) have gained massive momentum in the last few years, and achieved remarkable success in various tasks, either node, edge or graph classification. GNNs are well known for capturing semantic relations between one node and it's neighbors by propagating the neighbors messages to the current node, making those models well suited for embedding and classifying graphical data.

In this paper, we propose a new way to perform Active Learning using GNNs. Our scheme will work as follows, given tabular data, the first step is to construct a graph of all the data (labeled and unlabeled points) such that the nodes are the data points and edges are constructed based on similarity metric f, then the graph nodes will be embedded using GNN model which then embeddings of the labeled data will be passed on to train the classification head of the GNN, in parallel the unlabeled data will be converted into a graph and using centrality metrics ϕ, a subset of the nodes will be selected for labeling, which then the main classification model will be trained on, and we'll be back to the first step where we construct the graph on the unlabeled data.

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, δ be a hyper-parameter for similarity threshold, $\phi = \alpha \cdot \phi_{entropy} + \beta \cdot \phi_{density} + \gamma \cdot \phi_{centrality}$ is the uncertainty metric, 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

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

Algorithm 1 GNN Active Learning (Continued)

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 Experimentation and Evaluation

We used two methods as a baseline, the first one was the ActiveLearningPipeline with the Random uncertainty the second one was the same pipeline with Diversity sampling. We tested our GAL model on 5 datasets, **lab_2000** which is the dataset of lab 2 with 2000 sampled points from the available_pool, iris dataset, wine quality dataset, clustered and unclustered synthetic datasets, for evaluating the pipeline we used the accuracy evaluation function, although other robust and better evaluation could be used functions, but for the sake of this research the accuracy function was sufficient. while running the experiments we had to handle a trade-off between labeling effort which would be the graph construction, since calculating pairwise similarities, building the connections between the nodes and calculating centrality uncertainties are time consuming, therefore we had to work with relatively small datasets which had bad effects on the GNN performance.

4 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.

5 Discussion and Conclusions

Our proposed model, GAL, achieved results that were comparable to the baseline models, demonstrating promise in the use of graph-based Active Learning (AL) for tabular data. As this was our initial investigation in the field, we applied straightforward, reasonable methods to assess the feasibility of the approach without relying on advanced techniques. It's important to note that the GNN component in our model was trained on small datasets, which limited its ability to reach full potential. Despite this, GAL's comparable performance with baselines underscores the model's potential. Future work can focus on scaling GAL with larger datasets and incorporating more advanced methods to maximize its impact. This study sets a foundation for GAL's application to other domains and signals an encouraging direction for future exploration and refinement.

References

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

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

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*.