Graph Mining CSF426

Lab session 5 (evaluative)

Time: 5 pm – 7 pm

Date: 12-09-2024

Instructor IC – Vinti Agarwal

Instructions: All questions need to be answered. **You are required to write programs in jupyter notebook and submit .ipynb on canvas**. Please rename your solutions in format <ID-NAMELABNO>. For theoretical questions, you can type answers in the jupyter notebook itself. There is no need to create a separate text file. **[Total Marks =10]**

This lab exercise is based on **Correct and Smooth technique** explained in the research paper ***"Combining Label Propagation and Simple Models Out-performs Graph Neural Networks"*** (Link)

Objective: The objective of this lab session is to learn about process of label correction using error propagation.

Dataset: You are provided with cora dataset which contains information about 2708 research papers belonging to 7 categories of Machine learning. It is a citation network/graph where each publication/paper cites one or more other papers.

The dataset contains two files:

a. "Cora.content" – contains the description of research papers in format:

<PaperID> <1433 columns of word features> <class label> PaperID

– a unique identifier of each reseach paper.

Word features - each research paper is associated 1/0-valued word vector of length 1433 based on whether a word is present/absent in the corresponding paper.

Class label - contains the one category to which a paper belongs out of seven categories of : Case_Based, Genetic_Algorithms, Neural_Networks, Probabilistic_Methods, Reinforcement_Learning, Rule_Learning, Theory

b. "cora.cites" – contains the information about the citations. Each line describes a link in format:

<cited PaperID> <citing PaperID>

The first PaperID describes the ID of paper being cited and second PaperID is the ID of paper that contains the citation or that cited the first paper.

1. **Data loading:**
   a. Use ‘cora.conent’ file to extract the features and class information of each paper. Split the data into train/validation/test of ratio 60/20/20.
   b. Use ‘cora.cites’ file to create a graph.
2. **Base predictions:**
   a. First step is to make prediction using a plain linear model(**Logistic Regression)** to make base predictions on all the datapoints.
   b. The result will be base predictions $Z \in R^{n\times c}$ where c is the number of classes in dataset.
3. **Error propagation:** Use label spreading iteratively to propagate the errors in graph,
   a. Compute error matrix $\hat{E} \in R^{n\times c}, where$

   $$E_L = Z_L - Y_L, \qquad L - Training\ datapoint$$

   $$E_v = 0\,, \qquad v - validation\ datapoint$$

   $$E_U = 0\,, \qquad U - unlabeled/test\ datapoint$$

   And $E^{t+1} = (1-\alpha).\,E + \alpha.\,S.\,E^t\,, where\ E^0 = E\ and\ S = D^{\frac{-1}{2}}.A.\,D^{\frac{-1}{2}}$

   $$A - Adjacency\ matrix\ of\ data,\ D - Degree\ matrix\,,\ \alpha = [0.1{,}0.2{,}\ldots{,}0.9]$$

   Use validation data to find out the best alpha value.
   b. Corrected predictions,

   $$Z\_c = Z + \hat{E}$$
4. **Final prediction:** Use label spreading iteratively to make the final predictions
   a. Initialize matrix $G \in R^{n\times c}, such\ that$

   $$G_L = Y_L\ \ , G_v = Z_v,\ and\ \ G_U = Z\_c_U\ G^{t+1} = (1-\alpha).$$

   $G + \alpha.\,S.\,G^t, \quad \alpha = [0.1{,}0.2, \ldots{,}0.9],$
   b. Final classification, Y\` for a node $i \in U$ is $argmax_{j\in\{1,\ldots.,c\}} Y`_{ij}$
   c. Use validation data to find the best alpha value and report the accuracy results on best alpha.
5. Create a table of the accuracies achieved with base predictor, accuracy after error propagation, accuracy after final prediction with the accuracy achieved if we use Label Spreading algorithm (LSA) on the dataset directly.

NOTE: (you can use the code of your previous lab for LSA)

Method	Best alpha	Accuracy
Label propagation	-	
Plain linear		
Plain linear + C&S	Report best alpha for max accuracy	
Plain linear + autoscaling		
Linear +autoscaling		
MLP + autoscaling		