Final Project Proposal: Causality-Enriched Citation Prediction

Team: 啊對對對對隊
成員:
H24111057 姚博瀚、H24114089 陳峻霖、H24116049 莊秉宸

2024/12/15

(1) Motivations

Why is this project worth doing?

- Research citations are critical indicators of impact and influence in academia.
- Existing prediction methods are based on correlations, lacking interpretability and actionable insights.
- By incorporating causal discovery, we aim to build a model that identifies and leverages the direct drivers of citation counts.

Applications:

- Guide researchers in optimizing their work for better visibility and impact.

(2) Expected Dataset to be Crawled

Data Source: Google Scholar, PubMed, ScholarGPS, Web of Science, JCR

Metadata to be Crawled:

- Paper title, abstract, keywords, and journal name.
- Author information: affiliations, number of collaborators.
- Journal information: impact factor, open-access status.
- Citation counts and publication year.

(3) Problem Statement

Input:

- Sentence embeddings of abstracts, titles, and keywords, etc.
- Numerical features: collaboration size, journal impact factor, publication year up to now, etc.

Output:

- Predicted citation count for each paper.

Comparison to Existing Approaches:

- Current models focus on correlation, ignoring causal relationships.
- Our model incorporating causal discovery aims to improve interpretability and prediction robustness.

(4) Technical Challenges

Challenges:

- **Data Crawling:**
  - Handling Google Scholar rate limits and potential incomplete metadata.
- **Causal Discovery:**
  - Algorithms Selection (e.g., PC, FCI, LiNGAM, DAG-GNN, etc.)
- **Feature Integration:**
  - Deal with textual features and numerical features.
  - Applying causal penalty into loss function of regression model.

(5) Preliminary Methods

Feature Engineering:

- Use pre-trained BERT to extract sentence embeddings from abstracts and titles.
- Extract numerical features like collaboration size, journal impact factor.

Model Architecture:

- Neural network for embeddings and numerical features.
- Integrate causal discovery results into penalty during training.

Innovation:

- Leverage causal discovery algorithms (e.g.,PC, FCI, LiNGAM, DAG-GNN, etc.) to penalize features not causal to outcome.
- Predict citation counts while providing causal insights.

(5) Preliminary Methods

Algorithm 1 Causality-Weighted Citation Prediction

Require:

Sentence embeddings $\mathbf{E} \in R^{n \times d_e}$,
Numerical features $\mathbf{N} \in R^{n \times d_n}$,
Causal adjacency matrix $\mathbf{M} \in \{0,1\}^{(d_e+d_n)\times(d_e+d_n+1)}$,
True citation counts $\mathbf{Y} \in R^n$

Ensure:

Predicted citation counts $\hat{\mathbf{Y}} \in R^n$

Feature Extraction:
    Compute $\mathbf{E}$ using a pre-trained model (e.g., BERT).
    Normalize numerical features $\mathbf{N}$.
    Concatenate $\mathbf{E}$ and $\mathbf{N}$ to form combined features $\mathbf{X} \in R^{n\times(d_e+d_n)}$.
Causal Discovery:
    Apply a causal discovery algorithm (e.g., PC, FCI, DAG-GNN) to learn $\mathbf{M}$, the adjacency matrix with an extra column/row for outcome y.
    Extract causal mask $m_j = M_{j,y}$ for each feature j w.r.t. outcome y.
Model Training:
    Design a neural network $f_\theta(\mathbf{X})$ mapping $\mathbf{X_i}$ to $\hat{\mathbf{Y}}$.
    Let $\mathbf{w}$ be the weights mapping fused features to the output.
    Define the *causal-regularized loss* function:

$$\mathcal{L}(\mathbf{w}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}_{\text{MSE}} + \lambda \underbrace{\sum_{j=1}^{(d_e+d_n)}\left[1 - M_{j,y}\right]|w_j|}_{\text{Causal Penalty}}.$$

    Minimize $\mathcal{L}(\mathbf{w})$ w.r.t. $\mathbf{w}$ via backpropagation (e.g., Adam).
Prediction:
    Use the trained model f_θ to predict citation counts $\hat{\mathbf{Y}}$ for new inputs.

(6) Evaluation Plans

Training Settings:

- Split crawled data into train and test set.
- Hyperparameter tuning using Optuna for optimal performance.

Evaluation Metrics:

- Performace: RMSE, MAE, MSE, R^2.
- Causal graph quality: Conditional independence tests, expert validation.

Baselines:

- Standard models without penalty: XGBoost, Random Forest, etc.
- Ablation study to measure the impact of causal penalty.

(7) Expected Time Schedule

Week	Task
12/9-12/22	finalize crawling strategy, test crawling script.
12/9–12/22	preprocess crawled text and numerical features.
12/16-12/22	Perform causal discovery.
12/23-12/30	Build and train the model, validate results.
12/31-1/6	Write report and presentation slides.