IE 671 Web Mining Project Presentation

Team 7: Paper Importance Prediction

Roman Bogdanov, Nathanael Stelzner, Victoria Zevallos, Vitor Faria De Souza, Sharan Shyamsundar

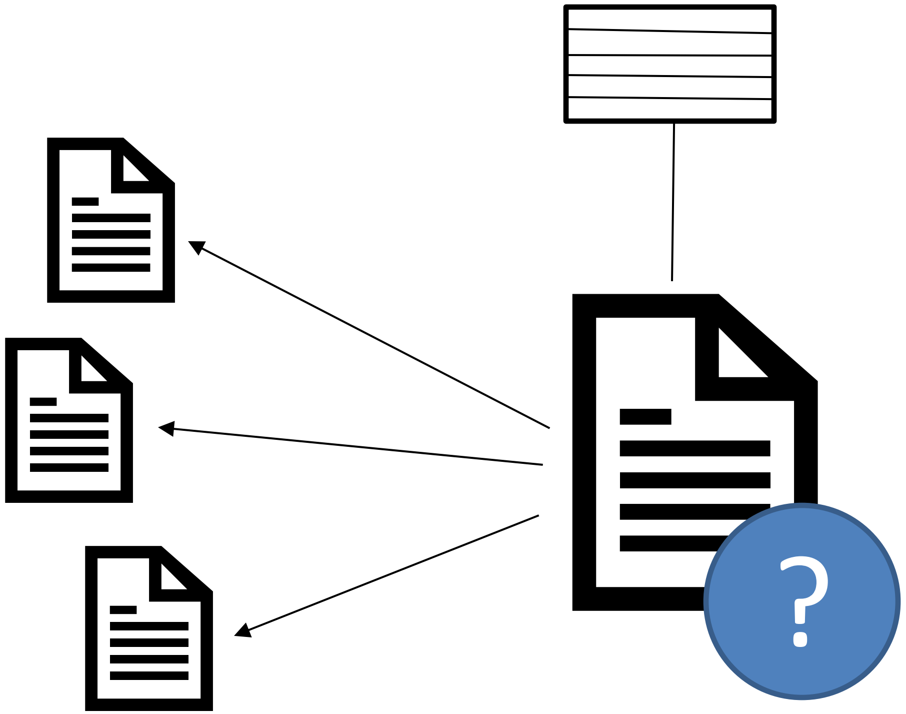
Presentation Outline

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

1. Introduction
2. Data and Preprocessing
    1. Data description
    2. Feature generation
    3. Target variable generation
3. Classical Models
4. Graph Neural Networks
5. Evaluation
6. Conclusion

Introduction

Task:



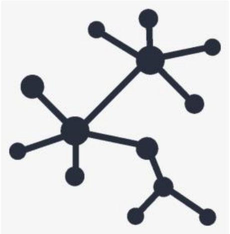
Questions:

1. Can ML methods predict future importance of scientific papers?
2. Which models are best suited for this task?
3. How do GNN models perform compared to the classical approaches?

Data and Preprocessing

Data Description

- Dataset provided by Stanford Network Analysis Platform (SNAP)
- 27770 scientific papers, each having a submission date and can cite other paper from the past
- Submission dates ranging from 1993 to 2003

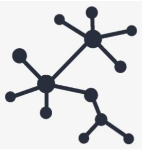


Each paper is a node and each citation is a directed edge	Meta information related to each paper
Nodes: 27770 Edges: 352807	Submission date, Submitter, comments, abstract, title and journal reference

Data and Preprocessing

Feature Generation

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



Network Data

- The graph was created using the text files provided.
- Calculated the in- and out-degrees of papers cited by the paper in question
- Some Features needed to be calculated with respect to time as future data should not influence the analysis

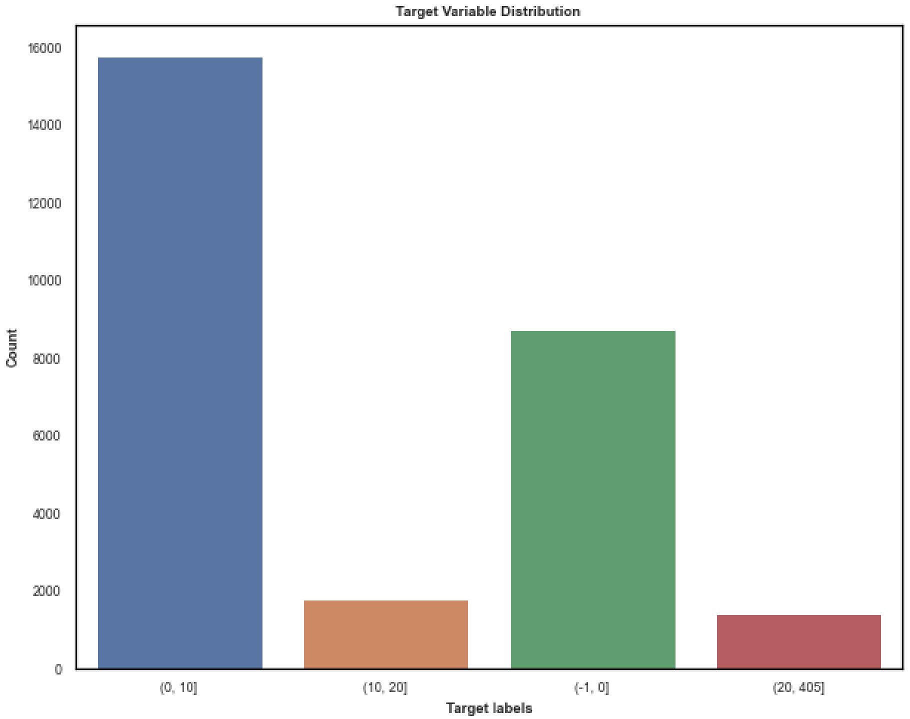
Metadata

- Used Regex to clean all text inconsistencies.
- Submitter name and email id is extracted from the submitter attribute
- From submission date, the first submission date and the number of revision was also computed.
- Number of pages and format of paper was derived using the comments feature
- Citations each submitter has received up to the date they publish a new paper is computed using the submitter details and citations data

Data and Preprocessing

Target Variable Generation

- We look at variety of targets for the number of citations
- Complication of highly unbalanced frequency distribution in case of 2 years
- The number of citations a each paper would receive in the first year of its publication
- Binned into 4 groups as the citations were too skewed



Classical Models

Standard data mining toolset - scikit-learn pipeline with:

- **Standard scaler**
- **Feature selection** based on ANOVA F-values (sklearn.feature_selection.SelectKBest)
- 5-fold cross-validation for **hyperparameter tuning**
- **Class weights** as inverse proportion of class sizes

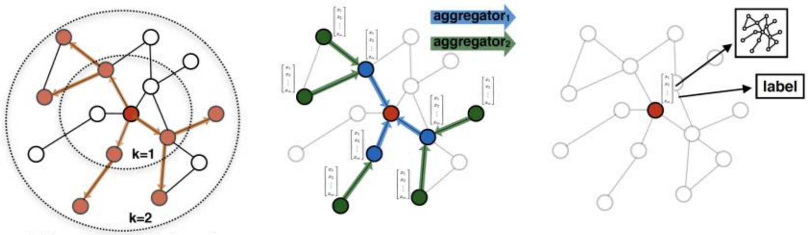
Applied models:

- Naive Bayes
- Multi-Layer Perceptron
- Decision Tree
- Random Forest
- Gradient Boosting
- Support Vector Machine

GNN

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Inductive learning
- Architecture: Graph Sage
  - 2 Graph Sage layers
    - 32 neurons
  - 2 Mean aggregator layers
  - 1 Dense layer (SoftMax)
    - 4 neurons



1. Sample neighborhood 2. Aggregate feature information from neighbors 3. Predict graph context and label using aggregated information

- 3 Model Configurations

Undirected graph	Directed graph	Directed graph
1L: 10 nodes 2L: 4 nodes	1L: 10 out-nodes 2L: 4 out-nodes	1L: 10 out-nodes 2L: 4 out-nodes
• Date Features	• Date Features	• Date Features • Metadata Features

- Training
  - Learning rate: 0.0001
  - 100 Epochs
  - Batch size: 50
  - Adam optimizer

[1] Hamilton, W., Ying, Z., Leskovec, J.: Inductive Representation Learning on Large Graphs. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)

Evaluation

- Test sets: papers submitted in the **end of dataset's timeline**
- Metric: **macro average F1 score** (imbalanced multiclass target)
- Baseline: **majority class** (1-10 citations in the first year = 53% of the test set, Macro avg F1 of 0.17)
- Result: **GNN outperformed** when using directed graph as input (0.49)

Classifiers	F1 *C0*	F1 *C1*	F1 *C2*	F1 *C3*	F1 Macro Avg
MULTI-LAYER PERCEPTRON	0.39	0.77	0.00	0.27	**0.36**
DECISION TREE	0.42	0.64	**0.18**	0.15	**0.34**
GRADIENT BOOSTING	0.37	**0.74**	0.06	0.10	**0.32**
RANDOM FOREST	0.24	0.76	0.0	0.0	**0.25**
SUPPORT VECTOR MACHINE	0.26	0.71	0.09	0.13	**0.30**
NAÏVE BAYES	**0.50**	0.67	0.16	**0.28**	**0.40**
GRAPHSAGEU (date)	0.00	**0.69**	0.00	0.00	**0.17**
GRAPHSAGED (date)	**0.99**	0.66	0.11	0.22	**0.49**
GRAPHSAGED (DATE+META)	0.91	0.59	**0.14**	**0.26**	**0.48**

Evaluation (+ Findings)

- **Metadata did not improve the GNN** (0.48-0.49)
- Undirected GNN predicted only the majority class for every paper
- All other models outperformed the baseline of 0.17
- **Naïve Bayes** performed best among classical approaches (0.40)

Classifiers	F1 *C0*	F1 *C1*	F1 *C2*	F1 *C3*	F1 Macro Avg
Multi-Layer Perceptron	0.39	0.77	0.00	0.27	**0.36**
Decision Tree	0.42	0.64	**0.18**	0.15	**0.34**
Gradient Boosting	0.37	**0.74**	0.06	0.10	**0.32**
Random Forest	0.24	0.76	0.0	0.0	**0.25**
Support Vector Machine	0.26	0.71	0.09	0.13	**0.30**
Naïve Bayes	**0.50**	0.67	0.16	**0.28**	**0.40**
GraphSageU (date)	0.00	**0.69**	0.00	0.00	**0.17**
GraphSageD (date)	**0.99**	0.66	0.11	0.22	**0.49**
GraphSageD (date+meta)	0.91	0.59	**0.14**	**0.26**	**0.48**

Conclusion

1. Paper Importance Prediction with ML methods possible
2. & 3. Directed GNNs best suited and better than classical ML methods

Future Direction:

- Use author(s) based features instead of submitter based features
- Use word embeddings from title and abstract
- Further feature generation

Thank you for your attention!

Questions?

Appendix

Feature importance examples

	feature_names	importance
2	recency_of_cited_papers_avg	0.162244
5	outdegrees_of_cited_papers_sum	0.137378
7	num_of_pages	0.134973
8	journal_counts	0.117096
3	max_time_difference_bw_cited_papers	0.101652
6	outdegrees_of_cited_papers_avg	0.081654
1	indegrees_of_cited_papers_avg	0.074447
0	indegrees_of_cited_papers_sum	0.070720
10	submitter_counts	0.059655
4	outdegree	0.045217
12	format_harvmac	0.007233
9	submitter_active	0.004555
11	journal_popularity	0.003175

Decision Tree Classifier

	feature_names	importance
4	outdegrees_of_cited_papers_sum	0.294121
2	max_time_difference_bw_cited_papers	0.164771
1	indegrees_of_cited_papers_avg	0.121482
5	outdegrees_of_cited_papers_avg	0.120719
0	indegrees_of_cited_papers_sum	0.095406
7	submitter_counts	0.083422
3	outdegree	0.071098
8	journal_popularity	0.025259
9	format_harvmac	0.020885
6	submitter_active	0.002836

Gradient Boosting Classifier

Hyperparameter optimization

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Classifier	Parameters optimized	Best parameter
Naive Bayes	Var_smoothing	1
Multi-Layer Perceptron	hidden_layer_sizes	6 (one hidden layer, 6 neurons)
Decision Tree Classifier	min_samples_split	5
	criterion	mse
	max_depth	7
	feature_selection_k	all
Random Forest Classifier	min_samples_leaf	5
	n_estimators	150
	max_depth	None
	criterion	entropy
	feature_selection_k	10

Hyperparameter optimization

Classifier	Parameters optimized	Best parameter
Gradient Boosting Classifier	min_samples_leaf	5
	max_depth	7
	criterion	mse
	feature_selection_k	10
Support Vector Machine	loss	hinge
	feature_selection_k	10

Detailed descriptions of all variables

variable	description
paper_id	Unique identifier of the paper
indegrees_of_cited_papers_sum	Sum of indegrees of papers cited by the paper in focus
indegrees_of_cited_papers_avg	Average of indegrees of papers cited by the paper in focus
recency_of_cited_papers_avg	Average recency of papers cited by the paper in focus (days?)
max_time_difference_bw_cited_papers	Maximum time difference between dates of publication of the cited papers
Outdegree	Outdegree of the paper in focus
outdegrees_of_cited_papers_sum	Sum of outdegrees of papers cited by the paper in focus
outdegrees_of_cited_papers_avg	Average of outdegrees of papers cited by the paper in focus
Submitter	The email id and Name of the submitter
submission_date	The date of submission of the paper and revision dates (if the paper is revised)
Title	Title of the paper
Authors	Authors of the paper
Comments	General comments regarding the paper like number of pages, format, number of figures etc.
report_no	Metadata
journal_ref	Metadata
Abstract	A paragraph to summarise the paper
submitter_email	Cleaned email id of submitter (if available)
submitter_name	Cleaned Name of submitter (if available)
Submitter_details	Cleaned unique identifier for each submitter (Mostly email, but if email is NA, then we take the name)
is_revised	Whether the paper was revised or not
times_revised	If revised, the number of times a paper was revised
first_submission_datetime	The first submission date with time
first_submission_date	The first submission date
num_of_pages	Number of pages of the paper (derived from comments)
Format	Format of the paper (derived from the comments)
journal_counts	Showing the number of times a particular journal occurs in the data
first_365_days	Numeric target variable
label	Target variable as category
label_name	Category description
submitter_counts	Number of papers submitted by the submitter
submitter_active	Submitter activity marker. 1 if submitter has submitted 5 or more papers, 0 if 4 or less
journal_popularity	Journal popularity marker. 1 if there are 13 or more papers in the database published in this journal, 0 if 12 or less
format_latex	Marker showing whether the paper was submitted in latex format. 1 if yes, 0 if no
format_revtex	Marker showing whether the paper was submitted in revtex format. 1 if yes, 0 if no
format_harvmac	Marker showing whether the paper was submitted in harvmac format. 1 if yes, 0 if no
format_plaintex	Marker showing whether the paper was submitted in plaintex format. 1 if yes, 0 if no
datedelta	Number of days passed between the paper being added to graph and the date of first publication. Note: multiple outliers exist.
Citations_till_date	For each submitter, the citations s/he has received till the date s/he publishes a new paper. (based on date added graph)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics