NGSA: NETWORK SCIENCE ANALYTICS CENTRALESUPELEC
Assignment 2

Team_Carbonara_without_creme_fraiche

Rebecca Erbanni
Centrale Supelec
Italy
rebecca.erbanni@student-cs.fr

Jithendra Sai Veeramaneni
Centrale Supelec
India
jithendrasai.veeramaneni@student-cs.fr

Jennifer Vial
Centrale Supelec
France
jennifer.vial@student-cs.fr

Sarah Bourial
Centrale Supelec
Morocco
sarah.bourial@student-cs.fr

Arnaud Cluzel
Centrale Supelec
France
arnaud.cluzel@student-cs.fr

1 INTRODUCTION

This assignment was aimed at completing a missing links prediction in a network of citation in scientific research articles. The network was under the form of a graph $G = (V, E)$: vertices represent the articles and edges between them indicate a citation of one paper by another.

2 DATA CLEANING

We had a total dataset 27 770 values.

2.1 Missing Values - Authors

We had a total of 4033 missing values for the column author. This was almost 1/5th of the dataset it was not feasible to dismiss or disregard these values so we decided to replace them with "Unknown". We also column divided the authors column into as many authors as each paper had to allow for a comparison of common authors.

2.2 Missing Values - Magazine

7472 missing values were found on the column magazine, almost 1/4th of our dataset. Similarly, we went against suppress these rows and replaced them by the value "missing". Magazines with incomplete names with likewise deemed not accurate and labelled as missing.

3 DATA ENGINEERING

The initial information contained within nodes included the paper title, publication year, author names and a short abstract. To accurately reconstruct the initial network, we made an attempt at enlarging our horizon of features using graph-theoretical, textual, and other information. We computed an additional total of 7 features in our approach. Since a big part of these new features engineering was based on values of authors or magazine that were missing, we created two variables to make our features engineering more accurate: mapping the magazines into categories and sub-categories. This was motivated by the observation that the papers were about scientific research but mostly on mathematics and physics, themselves divided into many branches of these two fields e.g. quantum physics, nuclear physics, astrophysics, algebra, geometry... etc. Further, this allowed us to make connections between papers depending on the category (and even sub-category) they belonged to i.e. belonging to the same (sub-)category means more chances of citation.

3.1 Number of overlapping words in a title

When two papers present the same title, or similar words thereof, there is a great chance of them both writing on the same topic and thus a chance of citation between the two scholars thereby creating a new connection between their papers. This feature measures the number of overlapping words in a paper title with another i.e. source paper and target citation paper. This variable is stored under `overlap_title`.

3.2 Temporal distance between the papers

The rationale behind this feature is that more contemporaneous research is usually more relevant, especially in the field of scientific research where innovation is on constant evolution. Thereby, two papers close in time seem more likely to be about the same research trend and therefore to cite one another. This variable is stored under `temp_diff`.

3.3 Number of Common Authors

Here we have a look at the intersection of authors across our nodes. Basically, papers written by the same scholar (so one who conducted the research of a source paper and a target paper) are very likely if not to be on the exact same subject, not belong to the same field or be an inspiration for a later work of the academic. Therefore, we expect to see citational links within the bibliography of a single author. This variable is stored under `comm_auth`.

3.4 Number of Overlapping words in description

This variable presents the same motivation as that of overlapping titles between two papers: if the abstract of two papers has a high number of overlapping (common) words, then chances are high both papers are on the same field and thus there is great potential for the authors to quote each otherâĂŹs work. To extract text features

from the papers abstracts, we used a TfidfVectorizer. This variable is stored under `overlap_description`.

3.5 Number of incoming edges

This feature computes the citation score of every paper i.e. the number of times it has been cited by other scholarly works. This variable is stored under `num_inc_edges`

3.6 No edges

This feature is basically about computing the number of papers that are isolated in citations i.e. are not cited and have not cited any of the papers in the dataset.

Features based on nodes neighborhoods

For a node x, let $\Gamma(x)$ denote the set of neighbors of x in G . A number of approaches are based on the idea that two nodes x and y are more likely to form a link in the future if their sets of neighbors $\Gamma(x)$ and $\Gamma(y)$ have large overlap; this follows the natural intuition that such nodes x and y represent authors with many colleagues in common, and hence are more likely to come into contact themselves i.e. an edge $< x, y >$ is more likely to form if edges $< x, z >$ and $< z, y >$ are already present for some z.

3.7 Common Neighbors

This is the most direct implementation of the above-mentionned idea for link prediction by defining score(x, y) := $|\Gamma(x) \cap \Gamma(y)|$, the number of neighbors that x and y have in common. This variable is stored under `comm_neighbors`.

3.8 Jaccard Index

The Jaccard coefficient is a commonly used similarity metric in information retrieval. It coefficient compares papers for two sets to see which papers are shared and which are distinct. It thus ensures the probability that both x and y have a feature f, for a randomly selected feature f that either x or y has. It is a measure of similarity for the two sets of data, with a range from 0 to 100 i.e. the higher the percentage, the more similar the two populations. The formula to find the Jaccard index is:
$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$
This gives a measure score (x,y) =: $\frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$
This variable is stored under `Jaccard`.

3.9 Shortest Path

Shortest path algorithms are a family of algorithms designed to solve the shortest path problem i.e. find the shortest path between only two given vertices s and t. This basic approach is to rank pairs of papers (x, y) by the length of their shortest path in graph G . Such a measure follows the notion that collaboration networks are âĂIJsmall worlds,âĂİ in which individuals are related through short chains. The shortest path between two scientists in wholly unrelated disciplines is often very short (and very tenuous).This variable is stored under `shortest_path`.

3.10 Adamic-Adar Index

Consider a related measure, in the context of deciding when two papers are strongly âĂIJrelated.âĂİ We compute features of the pages, and define the similarity between two pages to be the Adamic-Adar index:

$$\sum_{z=featuressharedbyx,y} \frac{1}{log(frequency(z))}$$

This refines the simple counting of common features by weighting rarer features more heavily. This suggests the measure score (x,y) := $\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log(\Gamma(z)}$.
This variable is stored under `Adamic`.

3.11 Cosine similarity

Cosine similarity is a similarity function that measures the angle between two vectors, and in our case - the angle between two documents. This is computed using the following formula:
similarity (x,y) = $\frac{X.Y}{||X||.||Y||} = \frac{\sum_j^n X_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$

Values range between -1 and 1, where -1 is perfectly dissimilar (unrelated papers, no common features) and 1 is perfectly similar (basically the same paper). This variable is stored under `cossim`.

3.12 Preferential Attachment

The basic premise is that the probability that a new edge involves node x is proportional to $|\Gamma(x)|$, the current number of neighbors of x. Further, literature suggests that the probability of citation of x and y is correlated with the product of the number of collaborators of x and y. This corresponds to the measure score(x, y) := $|\Gamma(x)||\Gamma(y)|$. This variable was stored under

4 MODEL

We first tried two algorithms: Random forest powered by Adaboost and Gridsearch which gave a Kaggle score of 0.851 and Long-Short-Term-Memory algorithm (LSTM) which gave a Kaggle score of 0.877.

Our highest score on Kaggle is 0.948. It was obtained using a supervised ensemble gradient technique for link prediction - Extreme Gradient Boosting (XgBoost), which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. XGBoost belongs to the group of widely used tree learning algorithms, which allow making prediction on an output variable based on a series of rules arranged in a tree-like structure. Our result did not seem surprising as gradient boosting methods are known in the literature to frequently be more performing than other algorithms.

5 EVALUATION

We tuned our model parameters manually after multiple attempts. All these features were also scaled before processing them into our model. Our output evaluation relied on the competition leaderboard score itself based on the F1 score itself delivered by Kaggle.