[F24] Data & Knowledge Representation
Extra Graph Approach (GA) assignment

Artem Matevosian

December 16, 2024

Milestone 1: small notes regarding competition environment

After reviewing the `FutureOfAIviaAI` repository, the following points deserve mentioning:

- The competition it is based off was held in 2021. That is, all the methods applied were already known in 2021. However, it is not guaranteed that newer methods and automatic representation learning will perform better: see the next point.
- The repository ranks the solutions by their performance decreasing [1, p. 1329]. End-to-end ML solutions (representation learning included) performed lower than ones using traditional features (solutions M7 and M8). On the other hand, both purely statistical/classical methods and their combinations with Graph Neural Networks (GNNs) scored higher than the baseline (M6).

Milestone 2: proposed technique and its justification

The approach will be leaning towards traditional graph node-level and link-level features with plain scoring and/or downstream classic prediction algorithms. The method might present some interest in the view of the fact that handcrafted features were competitive with representation learning and statistical techniques sometimes defeated pure ML in the original competition. The expected benefits of the approach are:

- low computational cost associated with statistical tests and some traditional node-level and link-level features over learned representations;
- extremely transparent and easily comprehensible theoretical motivation coming from graph and network theory, linear algebra, and statistics.

However, there is a drawback as well - most (if not all) traditional graph features default to zero or have no definition whenever the node in question is disjoint from the rest of the graph (degree zero), which might indeed be the case for some temporal states of the knowledge graph suggested in the assignment. It is still partially mitigated by the fact that the graph becomes more and more connected over time, having 1 connected component of size > 1 since before 2015 [1, Fig. 4].

Milestone 3: implementation specifics

3.1 Reading, storing and manipulating data

After using repository-defined snippets for reading from `SemanticGraph_delta_5_cutoff_25_minedge_3.pkl`, it was the most convenient to use `scipy.sparse` arrays/matrices representation - adjacency matrices of real-world graphs are usually sparse, and this dataset had not been an exception. Sparse representation allows for quick matrix and row/column operation along with much lower memory requirements, which made implementation of the feature generators possible.

3.2 Feature generation

The following are the features that were described in general in Milestone 2. They were generated for the whole set of unconnected node pairs. To speed up the computation, each generator worked in parallel in its own process. All the generators and multiprocessing code are located in `feature_gen.py`.

3.2.1 Preferential Attachment (PA)

This metric computes the product of the degrees of the two nodes in each pair:

$$\mathrm{PA}(u, v) = \deg(u) \cdot \deg(v)$$

Higher scores indicate a greater likelihood of connection based on the nodes' popularity.

3.2.2 Clustering Coefficient (CC)

The clustering coefficient of a node measures the density of connections among its neighbors:

$$\mathrm{CC}(u) = \frac{2 \cdot \#\ (\text{triangles through } u)}{\deg(u) \cdot (\deg(u) - 1)}$$

For a node pair, the link-level feature was computed as the product of their clustering coefficients:

$$\mathrm{CC\_Score}(u, v) = \mathrm{CC}(u) \cdot \mathrm{CC}(v)$$

3.2.3 Ruzicka Similarity (RZ)

Ruzicka similarity [2, S_5] can be considered a weighted version of Jaccard Index - it represents weighted intersection-over-union (IoU). For an adjacency matrix A, it is defined as:

$$\mathrm{RZ}(u, v) = \frac{\sum_w \min(A[u, w], A[v, w])}{\sum_w \max(A[u, w], A[v, w])}$$

One can see that for binary adjacency matrices the formula degenerates back to Jaccard Index.

3.2.4 Weighted Overlap Score (WOS)

This metric measures the overlap of neighbors weighted by degree. For adjacency matrix A [2, S_1]:

$$\mathrm{WOS}(u, v) = \frac{\sum_w \min(A[u, w], A[v, w])}{\min(\deg(u), \deg(v))}$$

3.2.5 Adamic-Adar Index (AA)

The Adamic-Adar index measures the strength of connection between two nodes by summing the inverse logarithms of their common neighbors' degrees:

$$\mathrm{AA}(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log(\deg(w))}$$

To avoid division by zero or $-\infty$, common neighbors with $deg(u) \leq 1$ were assigned zero terms (i.e. excluded from the sum).

Milestone 4: Evaluation

4.3 Feature-wise evaluation

For each feature, the whole set of currently unconnected pairs was sorted in monotonous (either ascending or descending) order of the feature values. Using this ranking, the ROC AUC score was computed to measure how well each single feature can serve as a predictor. The results are summarized in Table 1.

Graph/Network/Statistical Feature	ROC AUC Score
Preferential Attachment (PA)	0.9422
Clustering Coefficient (CC)	0.9380
Adamic-Adar Index (AA)	0.9495
Ruzicka Similarity (RZ)	0.7638
Weighted Overlap Score (WOS)	0.9288

Table 1: ROC AUC scores for individual features computed over whole test set.

4.4 Model-based Evaluation

Logistic Regression (LR) and Naive Bayes (NB) models were trained using the five features: PA, CC, AA, RZ, and WOS. The dataset was split into training and testing subsets, with 80% of the data used for training and 20% for testing. The models' predictions on the test set were used to compute the ROC AUC score in a similar manner to feature evaluation.

LR and NB models achieved ROC AUC scores of 0.9436 and 0.9459 respectively, meaning there is still potential of combining features via different methods for improved prediction (possibly, GNN variants supporting downstream features?)

Milestone 5: You are here!

I kindly ask you to consult `README.md` in `extra_model` folder to find out more, including running instructions.

References

[1] M. Krenn, L. Buffoni, B. Coutinho, *et al.*, "Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network," *Nature Machine Intelligence*, vol. 5, no. 11, pp. 1326–1335, Oct. 2023. DOI: 10.1038/s42256-023-00735-0.

[2] M. J. Warrens, "Inequalities between similarities for numerical data," *Journal of Classification*, vol. 33, no. 1, pp. 141–148, Apr. 2016. DOI: 10.1007/s00357-016-9200-z.