Using supervised learning techniques for entity relationships

[Extended Abstract]

Vipula Rawte
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180
rawtev@rpi.edu

Aparna Gupta
Lally School of Management
Rensselaer Polytechnic Institute
Troy, NY 12180
guptaa@rpi.edu

Mohammed J. Zaki
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180
zaki@cs.rpi.edu

ABSTRACT

Given different financial data resources, it is very challenging to relate entities across the various resources since each resource has its own way of describing the entities and relationships. We work on identifying such relationships using context and available scores, using mainly supervised machine learning techniques to build classifiers and predict new relationships or validate the existing ones based on the suitable measures of similarity.

CCS Concepts

•**Information systems** → *Content analysis and feature selection;*

Keywords

link prediction, classification, similarity score

1. INTRODUCTION

The Financial Entity Identification and Information Integration (FEIII) challenge tries to create and provide an interesting dataset in mainly the finance domain. It further focuses more on finding challenges and methods for solving them [1], [2]. This year the challenge aims to enhance a given "network" dataset by confirming the known relationships between two nodes of the network and further predicting unknown relationships. Given a network where the companies are linked to its competitors, suppliers, parents, subsidiaries, branches, etc., it is quite interesting to predict such relationships between a pair of company nodes. Link prediction in networks is important and useful since it can help with understanding the unknown association between two entities. Furthermore, given a large dataset, it becomes quite challenging to do the above task and so we use supervised learning algorithms to solve the challenge.

2. DATASET

The dataset used for this task was provided as a part of the FEIII Challenge 2018. It consists of the following datasets:

- 10-K reports
- Global legal Entity Identifier Foundation (GLEIF)
- Text-based Network Industry Classification (TNIC)
- Open Corporates
- Thomas Reuters Data Fusion (TRDF)

All the datasets contain a set of seed companies that are in the S&P 500 index, from North American Industry Classification System (NAICS) sectors 51 (Information) and 52 (Finance and Insurance).

The TNIC dataset consists of pairwise information which connects a financial entity (company's Central Index Key (CIK)) to its competitors. This pairwise information is called the **similarity score** and is in the range $[0.0, 1.0]$.

The TRDF dataset consists of information on relationships between two company nodes (CIKs). Following are the entity relationships in the given TRDF dataset:

- isImmediateParentOf
- isUltimateParentOf
- hasStrategicAlliance
- hasJointVenture
- isCompetitorOf
- isSupplierOf

Moreover, a subset of the TRDF ground truth dataset was provided for the scored task of link prediction.

DSMM'18, June 15, 2018, Houston, TX, USA
© 2018 ACM. ISBN 978-1-4503-5883-5/18/06…$15.00
DOI: https://doi.org/10.1145/3220547.3226044

3. METHOD

The scored task in this challenge is to predict the `isCompetitorOf` edge in the TRDF dataset. The predictions are to be made based on the similarity scores between two competitors in the TNIC dataset.

We use supervised learning classification algorithms like Support Vector Machines (SVM) and Random Forest Classifier for the prediction task.

We use the TRDF and TNIC data as our training data. Since there is only one feature in the training data, that is, similarity score, we construct more features based on it [3]. We use minimum, maximum and mean of the scores of the nodes adjacent to a given node. This leads to 7 features: min_first_node, min_second_node, max_first_node, max_second_node, mean_first_node, mean_second_node, and finally direct_similarity_score. Thus, our training data consists of TNIC direct scores, and our constructed scores along with the TRDF training data. The target variable is the predicate between two nodes. The value is set to 0 when no direct score exists between two entities.

The problem can therefore be formulated as a binary classification task, where 0 indicates the predicted edge is not an `isCompetitorOf` edge and 1 indicates it is an `isCompetitorOf` edge.

Table 1: Scored Challenge Task Results

	Precision (%)	**Recall (%)**	**F-score (%)**	**True Positive**	**False Positive**	**True Negative**	**False Negative**	**Total**
Ground Truth				211		4710		4921
	18.89	69.67	29.73	147	631	4709	64	4921

4. RESULTS

When we used SVMs on the ground truth dataset that consists of around 17K records, it unfortunately predicted all relationship instances as `isCompetitorOf`. On the other hand, a Random Forest Classifier (max_depth=2), gave much better results. The results when tested on the ground truth dataset are shown in Table 1. The challenge scores are quite low indicating that it might be difficult with the available training data. We certainly notice the precision-recall tradeoff here making it more recall-focused.

5. CONCLUSION

In this challenge we focused only on the edge prediction task using a Random Forest Classifier on the similarity score. We managed to get good scores using a relatively simple classifier. The main idea was to expand the similarity score based on the neighboring connections between nodes.

6. REFERENCES

[1] *DSMM'16: Proceedings of the Second International Workshop on Data Science for Macro-Modeling*, New York, NY, USA, 2016. ACM.

[2] L. Raschid, D. Burdick, M. Flood, J. Grant, J. Langsam, and I. Soboroff. Financial entity identification and information integration (FEIII) 2017 challenge: The report of the organizing committee. In *Proceedings of the 3rd International Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets, DSMM@SIGMOD 2017, Chicago, IL, USA, May 14, 2017*, pages 1:1–1:4, 2017.

[3] P. Sondhi. Feature construction methods: A survey. 2009.