Proposal Master Thesis

Investor Classification

University of Basel

Author:
Michael von Siebenthal

Supervisor:
Prof. Dr. Dietmar Maringer

March 23, 2021

Contents

1 Overview

The aim of this Master Thesis will be to analyze to what extent Graph Neural Networks (GNN) and its related methods (Kipf & Welling 2016, Hamilton et al. 2017, Veličković et al. 2017, Vaswani et al. 2017) can be applied for bank client classification. GNN methods have proven to be very successful for classification or prediction tasks among others. The advantage of these methods is that they can consider the network structure of a dataset (e.g. social network). This is a valuable feature if one for instance compares it to the capabilities of Convolutional Neural Networks (CNN) which can "only" work with grid structures (e.g. image classification). GNN methods are currently widely used in recommendation systems for social media (e.g. Pinterest, Ying et al. (2018)) and could be similarly beneficial for banks when classifying clients. Credit Scoring for instance appears to be an application in which GNN perform very well as shown by Sukharev et al. (2020).

Another interesting application and the focus for this thesis would be to classify bank clients according to their investment preference (e.g. which type of products should be advertised to which client?). This would be especially useful in the retail banking segment where advisors typically cannot know their clients personally due to the large number of clients being serviced. Investor classification is the intended main focus of the Master Thesis.

This topic faces many different hurdles due to the low availability and mostly absence of available bank client data. To the extent possible, appropriate data sets will be retrieved (thus far 1 dataset found). The main difficulty however is to find a dataset which both includes attribute/feature data and the network structure of the customer data. For this reason, mostly methods to create synthetic data will be used to created the dataset for the subsequent simulation/testing.

In the subsequent sections the provisional outline of the intended Master Thesis is briefly presented.

2 Introduction

This section will provide an overview of network analysis and its applications (e.g. social network analysis, protein folding (AlphaFold)). In addition, applications of GNN that are specifically relevant for business & economics are presented which are among others:

- Information flow through organizational charts
- Relationships / interdependencies between banks

- Compliance tasks
- Client classification

This section will constitute a smaller portion of the Master Thesis and is meant to provide an introduction to the topic of network analysis, why it is relevant to business and economics and describe what the aim of the Master thesis is.

3 Theory

The following subsections will introduce the theory used in the Master Thesis.

3.1 Graph Theory

In this section a brief introduction to the most important concepts of graph theory will be given. This section will include the relevant concepts which in subsequent chapters will be used to analyze the network for the client classification task. The theory in this sections will be based on the book "Networks: An Introduction" by Mark Newman (2010) and the Stanford open lectures by Prof. Jure Leskovec.

3.2 Graph Neural Networks

This section will introduce the relevant theory required for Graph Neural Networks. In a first step node embedding strategies such as DeepWalk (Perozzi et al. 2014) and node2vec (Grover & Leskovec 2016) are introduced which are useful methods for embedding networks into lower dimensional space. Afterwards it is shown how GNN provide a more general and scalable approach as well as how they can be used for classification tasks (e.g. Kipf & Welling (2016)).

3.3 Synthetic Graph Data

This section will provide an overview of the theory and approach for the synthetic data generation methods. The method applied will be primarily based on the "Multiplicative Attribute Graph" (MAG) model (Kim & Leskovec 2012). This method allows to create completely synthetic networks incl. node attributes. It also allows for known node embeddings to be used for the network generation process. The benefit of this method lies in that it can create networks which resemble real-life network structures, such as social networks.

As mentioned in section 1, bank client data is most likely not going to be available at a useful scale or form. For this reason synthetic data generation methods will be used. For this reason the following staged approach will most likely be used:

1. Create or use an existing dataset such as the bank dataset by Moro et al. (2014)
2. Use the MAG model to create network connections between nodes
3. If necessary grow the network using the MAG model or if available alternative methods.

This methods will be used for synthetic, existing and self created (survey) datasets in order to create the network connections and to derive a sufficiently large data set for the subsequent prediction task.

4 Research Question

As shown in the previous section, due to limited data access it will have to be created in large part synthetically. The situation in which the feature data of the clients/nodes is known while the network connections are unknown is of particular interest for this Master Thesis. From a research perspective it is typically relatively feasible to attain feature data of clients related to a topic of interest via surveys among other possibilities. It is however very difficult to in addition also capture the associated network connections in a sensible way. While banks do have a unique advantage in this regard in that they can build network data based on client transactions (e.g. payments), access to such data is rarely granted. Rare exceptions are for instance the article by Sukharev et al. (2020), where they received access to anonymized bank data for developing a credit scoring system.

This leads to the research question of this Master Thesis:

> *To what extent is partially synthetic network data predictive for bank client classification using Graph Neural Networks?*

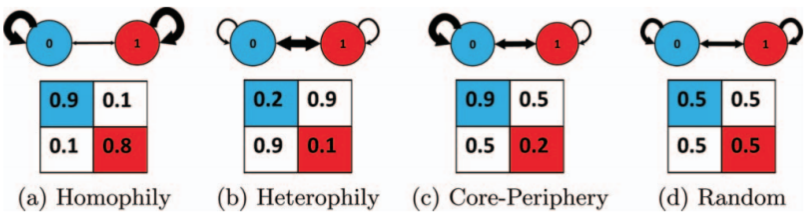
The partial synthetic network data refers to datasets where the feature data is known, while the network connections are synthetically created. Most likely the dataset by Moro et al. (2014) will be used to answer this question. Specifically the accuracy of GNNs will be compared to the results of the non-GNN method used by Moro et al. (2014).

5 Application

An application of GNN which will focus on client classification (investor type) will be presented in this section. As mentioned in section 1, advisors servicing retail clients cannot personally know their clients due to the large volume of clients. For this reason often only generic offers are advertised to this client group. For instance an active investor interested in trading stocks or bonds will most likely be advertised mutual funds. This leads to a mismatch which could be resolved through GNN. In order to test the suitability of GNN, a survey will be conducted with at least 100 participants (preferably many more). The survey will ask participants a list of attribute questions as well as their investor type. The investor types and the attributes will be selected based on existing literature and expert interviews. I have worked for over 10 years as an advisor for a large Swiss bank and have good contacts to conduct expert interviews to accompany the extant literature regarding investor type classification. Once the survey is completed, a network structure using the GAM method will be added to the dataset. The analysis of the application will perform an investor classification task using the collected survey data. It will be interesting to see to what extent the client classification task can be performed using the small survey dataset. As the collected survey data will most likely be too small for performing a GNN analysis, the network will be grown using the MAG method or an alternative method. If the MAG method is used, the network will be created synthetically by using the mean of the survey dataset with regard to the random assignment of the features using Bernoulli trials.

5.1 Multiplicative Attribute Graph

As described in the previous section, the network will be created using the Multiplicative Attribute Graph method. A crucial ingredient for the graph creation are the affinity matrices. Depending on the link-affinity probability, the resulting network structure can differ significantly. The link-affinity matrices provide for an area where the probabilities can on the one hand be set based on theory, they however also allow for simulating different scenarios. The synthetic graph creation is best described using the illustration provided by the authors (Kim & Leskovec 2012):



(a) Homophily (b) Heterophily (c) Core-Periphery (d) Random

Figure 1. Node-attribute link affinities. Four different types of link affinity of a particular attribute. Circles represent nodes with attribute value 0 and 1, and the width of the arrows corresponds to the affinity of link formation between the two groups. Each bottom figure indicates its corresponding link-affinity matrix (color figure available online).

Figure 1: Link-Affinity Matrix
Kim & Leskovec (2012)

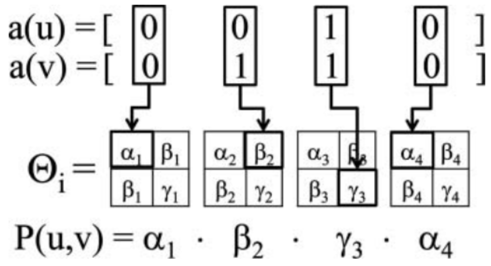


Figure 2. Schematic representation of the multiplicative attribute graphs (MAG) model. Given a pair of nodes u and v with the corresponding binary attribute vectors $a(u)$ and $a(v)$, the probability of an edge $P[u, v]$ is the product over the entries of attribute link-affinity matrices Θ_i, where values of $a_i(u)$ and $a_i(v)$ "select" appropriate entries (row/column) of Θ_i. Note that this particular model represents an undirected graph by making each link-affinity matrix Θ_i symmetric. However, the MAG model generally represents directed graphs.

Figure 2: Multiplicative Graph Model
Kim & Leskovec (2012)

6 Limitations/Risks

The Master Thesis has some limitations and risks regarding myself as the author and the application:

1. The topic of GNN is a relatively recent topic and at the limit of my personal programming capabilities. There are fortunately packages such as Pytorch Geometric among others which can compensate for some of my programming deficiencies. Nevertheless, even with the support of such packages, the programming skills required is ambitious given the typical programming experience of an economics student. I am however confident, that I can successfully complete the project.

2. The GAM method has shown to create realistic real world networks using attribute data. Nevertheless, this method might fail for analyzing investor types and would negatively impact the accuracy of the GNN.

3. The survey questionnaire must capture relevant attributes so that the GNN can perform as intended. Failure in this area would negatively impact the results. In addition, acquiring sufficient survey participants is a difficult challenge.

4. If the GNN does not provide good results, this will put in question the application for which the survey will be conducted. In this case reasons for the failure will be discussed. Further, an alternative classification method will be used for the classification task.

5. The attractiveness of the Master Thesis relies on the success of the GNN methods chosen. Nevertheless, a failure would point to the limitations of synthetic network

creation. Possible reasons for the failure would be researched and discussed which in itself would yield useful insights for future research. Nevertheless the aim of the Master Thesis will be to provide a working GNN model.

References

Grover, A. & Leskovec, J. (2016), node2vec: Scalable feature learning for networks, *in* 'Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 855–864.
URL: *https://doi.org/10.1145/2939672.2939754*

Hamilton, W. L., Ying, R. & Leskovec, J. (2017), 'Inductive representation learning on large graphs', *arXiv preprint arXiv:1706.02216* .
URL: *https://arxiv.org/abs/1706.02216*

Kim, M. & Leskovec, J. (2012), 'Multiplicative attribute graph model of real-world networks', *Internet mathematics* **8**(1-2), 113–160.
URL: *https://doi.org/10.1080/15427951.2012.625257*

Kipf, T. N. & Welling, M. (2016), 'Semi-supervised classification with graph convolutional networks', *arXiv preprint arXiv:1609.02907* .
URL: *https://arxiv.org/abs/1609.02907*

Moro, S., Cortez, P. & Rita, P. (2014), 'A data-driven approach to predict the success of bank telemarketing', *Decision Support Systems* **62**, 22–31.
URL: *https://doi.org/10.1016/j.dss.2014.03.001*

Newman, M. (2010), *Networks: An Introduction*, Oxford University Press, Inc.

Perozzi, B., Al-Rfou, R. & Skiena, S. (2014), Deepwalk: Online learning of social representations, *in* 'Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 701–710.
URL: *https://doi.org/10.1145/2623330.2623732*

Sukharev, I., Shumovskaia, V., Fedyanin, K., Panov, M. & Berestnev, D. (2020), 'Ews-gcn: Edge weight-shared graph convolutional network for transactional banking data', *arXiv preprint arXiv:2009.14588* .
URL: *https://arxiv.org/abs/2009.14588*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), 'Attention is all you need', *arXiv preprint arXiv:1706.03762* .
URL: *https://arxiv.org/abs/1706.03762*

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. & Bengio, Y. (2017), 'Graph attention networks', *arXiv preprint arXiv:1710.10903* .
URL: *https://arxiv.org/abs/1710.10903*

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L. & Leskovec, J. (2018), Graph convolutional neural networks for web-scale recommender systems, *in* 'Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', pp. 974–983.
URL: *https://doi.org/10.1145/3219819.3219890*