Graph Machine Learning

Studiengang: MAS Data Science

Der Einsatz von Graph Datenbanken zusammen mit Machine Learning Algorithmen bietet im Bereich der Datenwertschöpfung neue Möglichkeiten. Die Master-Thesis zeigt, dass mit Hilfe von modernen Graph Machine Learning Algorithmen wie Node2Vec oder GraphSAGE Vorhersagen für fehlende Verbindungen gemacht werden können, die mit regelbasierten Ansätzen nicht möglich sind.

Ausgangslage

Der «Enterprise Data Catalog» (EDC) der Mobiliar Versicherungsgesellschaft importiert Metadaten von verschieden Datenquellen und führt die Datensätze in einer Graph Datenbank zusammen. Eine Aufgabe ist die Verknüpfung von Namen aus Datensammlungen mit den effektiven Mitarbeitenden der Stammdaten. Aktuell wird diese Verknüpfung regelbasiert gemacht. Da die Namen in beliebiger Form erfasst werden, ist eine Zuordnung aber nicht immer möglich.

Zielsetzung

Mit der Master-Thesis soll das Wissen über Graphen in Kombination mit dem Einsatz von Machine Learning Algorithmen aufgebaut und mit praktischen Experimenten nachvollzogen werden.

Darauf basierend soll die Verknüpfungsproblematik der EDC-Anwendung mit diversen Link-Prediction Techniken untersucht und verglichen werden.

Resultate

Es hat sich gezeigt, dass die klassischen «Similarity based Link Prediction»-Ansätze für die Vorhersage der gegebenen Datenkonstellation des EDC nicht geeignet sind. Moderne «Learning based Link Prediction»-Algorithmen wie Node2Vec oder GraphSAGE hingegen erzielten gute bis sehr gute Resultate. Damit können Vorhersagen für fehlende Verbindungen gemacht werden, die mit dem aktuellen regelbasierten Ansatz nicht möglich sind.

Auch wenn die Versuche erst mit synthetischen Testdaten durchgeführt wurden, zeigt die Master-Thesis gut, was diese Algorithmen bieten und wie zum Beispiel mit der Modellierung von Node Features spezifische Fragestellungen adressiert werden können. Für einen Einsatz in der Praxis sind sicher noch Tests mit realen Daten sinnvoll. Zudem gibt es weitere interessante Algorithmen, bei denen sich eine Untersuchung lohnt.

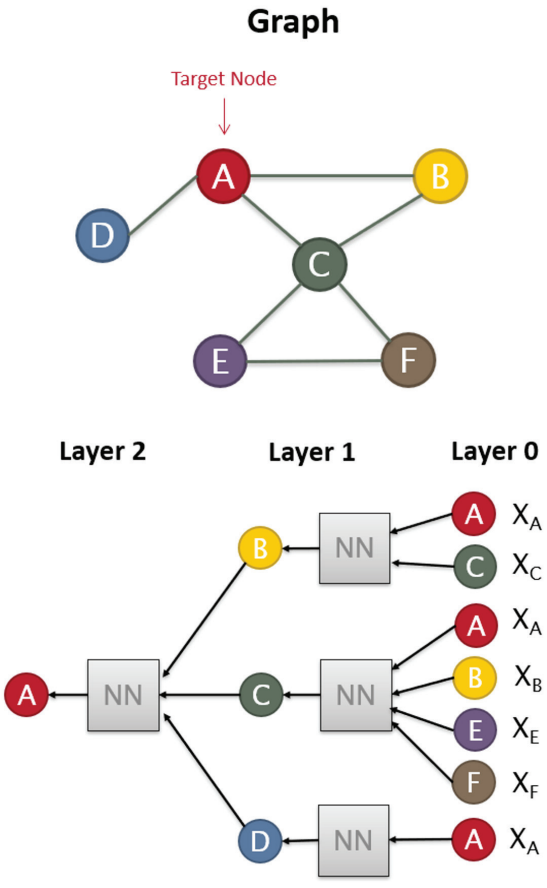
Fazit

Die EDC-Graph-Datenbank der Mobiliar zeigt schon jetzt sehr eindrücklich, wie schnell verschiedene Datenquellen miteinander verknüpft und ausgewertet werden können. In Kombination mit Graph Machine Learning ergeben sich neue Möglichkeiten, weiche in Zukunft sicher noch an Bedeutung gewinnen werden.

Quellcode

https://github.com/surfmachine/gml

Thomas Iten
tom@globalfootprint.ch



Graph-Convolutional-Network Neighborhood Aggregation