Graph Machine Learning

Explorative Arbeit im Bereich Link Prediction

Antrag für Zulassung Master Thesis
BFH, Technik und Informatik, Weiterbildung

1*) Angaben Autor/Autorin:

Autor/Autorin:	Thomas Iten
Ausbildung/Höchster Abschluss:	Diplom Ingenieur FH Elektrotechnik Nachdiplom Softwareingenieur HTL-NDS Nachdiplom Wirtschaftsingenieur FH-STV
Firma/Ort:	Die Mobiliar, Bern
Studiengang:	MAS Data Science

2*) Eingangskompetenzen:

1. CAS/Semester:	CAS Datenanalyse / HS19
2. CAS/Semester:	CAS Practical Machine Learning / FS20
3. CAS/Semester:	CAS Data Visualization / HS20
4. CAS/Semester:	CAS Artificial Intelligence / HS21

	Ja	Nein
Obligatorische WAW-Blöcke 1-3 besucht oder geplant	☒	☐

Ergänzende Angaben/Bemerkungen: z.B. externe CAS, vorgezogene Master Thesis, Zulassung über DAS oder individuelle Vereinbarung, Dispensation WAW usw.	

	Ja**	Nein
Relevante Vorarbeiten (LC/Semesterarbeit)	☐	☒

* Bei Gruppenarbeiten Pos. 1) und 2) pro Person ausfüllen.
** Wenn ja, bitte beilegen.

3) Angaben zur Master-Thesis:

Semester:	FS22
Themensponsor (Firma):	Die Mobiliar, Bern
Themensponsor (Betreuungsperson):	Matthias Brändle Verantwortlicher Datenstrie
Art der Arbeit/ Referenzrahmen	Explorative Arbeit

	Ja***	Nein
Vertraulichkeitsvereinbarung/NDA	☐	☒

Ort/Datum:	Bern, 8. Januar 2022

*** Wenn ja, bitte beilegen.

1 Ausgangslage und Motivation

Mit Graphen lassen sich komplexe Strukturen wie physikalische Systeme, Molekularstrukturen, Verkehrsnetze oder soziale Netzwerke abbilden. Neben der Graph Struktur beinhalten die Knoten und Verbindungslinien der Graphen ebenfalls wertvolle Eigenschaften. Für die Verarbeitung solcher Informationen kommen unter anderen Machine Learning (ML) Algorithmen zum Einsatz. Diese sind in der Lage Informationsmerkmale mit Graph Strukturen zu kombinieren, und davon zu lernen. Sie helfen uns, die Informationen der Graphen zu verstehen, Strukturen und Merkmale besser zu modellieren, um darauf basierend Vorhersagen zu treffen. Damit können reale Anwendungen wie das Erlernen von molekularen Fingerabdrücken, die Kontrolle von Verkehrsströmen oder Empfehlungen in sozialen Netzwerken realisiert werden.

Das Potential von Graph Anwendungen ist sehr gross, so dass diese in der Industrie immer mehr zum Einsatz kommen. Mit den beiden CAS «Practical Machine Learning» und «Artificial Intelligence» konnte ich erste Erfahrungen mit Machine Learning und im speziellen mit neuronalen Netzen sammeln. Die Kombination von Graphen mit ML ist spannend und fasziniert mich persönlich sehr. Gerne würde ich das Thema im Rahmen der Master-Thesis intensiver erkunden und praktische Erfahrungen sammeln.

2 Zielsetzung / Fragestellung

Mit der Master-Thesis soll das Wissen über Graphen in Kombination mit dem Einsatz von ML Algorithmen aufgebaut und mit praktischen Experimenten nachvollzogen werden. Dazu werden öffentliche Daten ausgewählt und eingesetzt, so dass die Erkenntnisse und Experimente für alle Interessierten im Internet frei zugänglich sind. Ziel ist es, mit dem so erlangten Wissen abschätzen zu können, für welche Geschäftsanwendungen die Technologie in Zukunft nutzbringend eingesetzt werden kann.

Darauf aufbauend soll eine Fragestellung im Kontext der Mobiliar Versicherung vertieft untersucht werden. Mit dem «Enterprise Data Catalog» (EDC) stellt die Mobiliar eine Graph basierte Anwendung zur Verfügung. Diese importiert die Metadaten von verschieden Datenquellen (Datenbanken, Protokollen, Mitarbeiter Stammdaten, etc.) und führt die Datensätze in einem Graphen zusammen. Eine Aufgabe ist dabei die Verknüpfung von Teilnehmernamen der Sitzungsprotokolle mit den effektiven Personen aus den Stammdaten. Aktuell wird diese Verknüpfung regelbasiert gemacht. Da die Protokollnamen in beliebiger Form erfasst werden ist eine Zuordnung aber nicht immer möglich.

Die Vorhersage von fehlenden Verbindungen (Link Prediction) ist eine häufige Aufgabe beim Arbeiten mit Graphen. Die zu untersuchenden Fragestellung lautet also: Wie können mit Link Prediction Techniken die Daten von verschiedenen Datenquellen effizient miteinander verknüpft werden? Konkret soll dazu in einer Fallstudie die erwähnte Verknüpfungsproblematik der EDC Anwendung mit diversen Link Prediction Techniken (wie Ressource Allocation, Jaccard Coefficient, Common Neigbor oder Supervised Classification) untersucht und verglichen werden.

Dabei sollen Fragen, wie folgende, beantwortet und dokumentiert werden:

- Wie kann eine geeignete Testumgebung für die Experimente aufgebaut werden?
- Wie werden relevante Testdaten der EDC Anwendung extrahiert und anonymisiert?
- Welche Python Bibliotheken für Graph ML Aufgaben gibt es?
- Welche Kategorien (Taxonomie) von Link Prediction Techniken gibt es?
- Welche Link Prediction Techniken werden in der Fallstudie untersucht, welche Python Bibliotheken kommen zum Einsatz?
- Wie werden die Ergebnisse der Untersuchungen bewertet und miteinander verglichen?
- Welche der untersuchten Varianten eignen sich für die Verknüpfung der Personendaten?

3 Methoden

Für die explorative Arbeit wird eine agile Vorgehensweise nach SCRUM gewählt. Die Anforderungen werden in Form von Stories erfasst, mit dem Themensponsor abgeglichen und iterativ umgesetzt. Sobald eine Story umgesetzt ist, werden die Resultate den Betreuern und dem Themensponsor präsentiert und die nächsten Schritte geplant. Die Durchführung der Experimente orientiert sich am Experimentierzyklus aus dem Lehrplan 21 (Klett & Balmer, 2019).

In der ersten Phase wird das Grundlagenwissen zum Thema erarbeitet und jeweils mit praktischen Experimenten vertieft und verifiziert. Parallel dazu erfolgt der Aufbau der notwendigen Entwicklungs- und Testumgebung.

In der zweiten Phase werden in der Fallstudie verschiedene Link Prediction Techniken zur Verknüpfung von Personendaten erkundet und mit Experimenten getestet. Als Testdaten werden entweder die relevanten Einträge aus der Mobiliar EDC Anwendung extrahiert und anonymisiert oder öffentliche Daten mit analogen Eigenschaften verwendet.

Die Zwischenergebnisse und Erkenntnisse der Arbeiten werden laufend festgehalten und dokumentiert. In der dritten Phase folgt die Finalisierung des Berichts.

4 Abgrenzung

Der Fokus der Arbeit liegt im Erkunden und Vergleichen der verschiedenen Link Prediction Techniken zur Verknüpfung der genannten Personendaten. Die technische Integration von geeigneten Algorithmen in die EDC Anwendung ist nicht Bestandteil der Master-Thesis.

In den Experimenten sollen keine heiklen Daten zum Einsatz kommen, so dass alle Experimente öffentlich zugänglich gemacht werden können.

5 Rahmenbedingungen

- Sind die benötigten Informationen vorhanden? Wo finde ich die Informationen?

  Ja, es gibt öffentliche Artikel und Arbeiten zum Thema sowie entsprechende Frameworks zum Umsetzen von Experimenten.

- Welches sind die Risiken/Widerstände? Was kann ich dagegen tun?

  Das Thema ist modern und in stetem Wandel. Es gibt viele verschiedene Techniken zur Link Prediction und eine aktuelle Gesamtübersicht ist schwer zu finden. Die Master-Thesis fokussiert sich auf die aufgeführten Algorithmen. Stellt sich im Laufe der Arbeit heraus, dass andere Algorithmen erfolgsversprechender sind, wird die Zielsetzung in Absprache mit den Experten angepasst.

- Gibt es fördernde Faktoren/Chancen? Wie nutze ich diese?

  In der Mobiliar wird die gezielte Nutzung von Machine Learning und Artificial Intelligence gefördert. Mit der EDC Anwendung ist zudem eine Graphen Datenbank im Einsatz. Entsprechend hat es erfahrene Mitarbeiter, die für einen produktiven Austausch zur Verfügung stehen.

- Benötigte Infrastruktur, Material zur Durchführung eines Experiments, Fachexperten?

  Die konkreten Anforderungen sind noch nicht bekannt. Wenn möglich sollen die Experimente lokal auf einem PC oder mit Google Colab ausgeführt werden können. Falls die Kapazitäten nicht ausreichen, besteht die Möglichkeit eine ML Umgebung mit Azure (Microsoft Cloud) aufzusetzen und zu verwenden.

- Abhängigkeit von den Resultaten oder Ergebnissen vorangehender Projekte, anderer Teilprojekte im Unternehmen?

  Eine direkte Abhängigkeit welche für die Erarbeitung der Master-Thesis zwingend erforderlich ist gibt es nicht.

6 Deliverables / Resultate

Basierend auf den Zielen und der Fragestellung sollen folgende Ergebnisse erarbeitet werden:

1. Wissensaufbau und Dokumentation der Grundlagen von Graph Machine Learning, so dass abgeschätzt werden kann, für welche Geschäftsanwendungen die Technologie in Zukunft nutzbringend eingesetzt werden kann.

2. Fallstudie mit Vergleich verschiedener Link Prediction Ansätze zur Verknüpfung von Personendaten der EDC Anwendung.

3. Experimente mit Source Code (Python und Jupiter Notebooks) zum Nachvollziehen der einzelnen Methoden und Verfahren.

4. Schlussbericht

7 Anhang

Die sieben Phasen des Experimentierzyklus:

Klett, & Balmer. (2019). *Klett*. Von Die sieben Phasen des Experimentierzyklus: https://downloads.klett.ch/uploads/images/textbooks/experimentierzyklus-didaktik-prisma-klett-und-balmer.jpg abgerufen