DAIKIRI

Erklärbare diagnostische KI für industrielle Daten

Titel:

Herausforderungen im Umgang mit KI für KMU am Beispiel der Erklärbarkeit

Partner:

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

pmOne USU elevait DICE

Sustainable AI makes the difference

1. Abstrakt

Künstliche Intelligenz (KI) hat in den zurückliegenden Jahren erstaunliche Fortschritte gemacht und sich zu einer Schlüsseltechnologie entwickelt. Jedoch hat KI die Wirtschaft weitaus weniger durchdrungen, als erhofft. In diesem Whitepaper wird auf die Hintergründe eingegangen und die Herausforderungen bzgl. der fehlenden Erklärbarkeit näher beleuchtet. Eine neue, hoch-innovative Möglichkeit für die Schaffung erklärbarer KI-Modelle wurde im Forschungsprojekt DAIKIRI entwickelt, dessen Konzept im Überblick erläutert wird.

2. Notwendige Digitalisierung, Einsatz von KI, generelle Einführung zum Thema KI und deren Herausforderungen

Der Mittelstand ist ein wesentlicher Innovations-, Technologie- und Wirtschaftsmotor Deutschlands. Er ist schnell und reagiert auch in Krisenzeiten flexibel auf neue Marktsituationen. Mehr als 99% der deutschen Unternehmen sind KMUs, die nahezu 60% aller Arbeitsplätze in Deutschland stellen und ca. 55% der Nettowertschöpfung tragen[1]. Insgesamt machen deutsche KMUs einen jährlichen Umsatz von ca. 2 Mrd. EUR[2]. Aber sie stehen auch vor der riesigen Herausforderung der Digitalen Transformation, bei der sie sich als Nachzügler sehen[3]. Dies bedeutet, sie müssen ihre internen Wertschöpfungsprozesse optimieren, neue Wege zur Kundenansprache und -aktivierung nutzen, oder neue Daten-/KI-getriebene Produkte und Dienstleistungen schaffen[4]. Um diese und weitere Herausforderungen zu lösen, wird **künstliche Intelligenz** (KI) als eine wesentliche Technologie gesehen[5] [6]. So betrug schon 2019 das Umsatzvolumen, dass in Deutschland von KI beeinflusst wird, ~220 Mrd. €[7] und der Markt wird in den verschiedenen Anwendungsfällen für KI-Software, wie etwa dem Wissensmanagement, weltweit stärker wachsen[8]. Jedoch ist der KI-Einsatz sehr viel begrenzter, als es vielseitig dargestellt wird[9]. Die Gartner-Studie ergab[10], dass Unternehmen häufig mit KI experimentieren, aber Schwierigkeiten haben, die Technologie in ihre Standardabläufe einzubinden. Die Hintergründe dazu sind verschieden, wesentliche Faktoren werden jedoch nachstehend kurz erläutert[11]:

KI als Projekt: Kaum eine Unternehmenssoftware beinhaltet heute KI als Produkt-Bausteine für verschiedene Anwendungsfälle. Dies hat verschiedene Gründe, wie etwa, dass die Software-Architekturen durch ihr hohes Alter wenig modular, kaum agnostisch bzgl. der Technologien oder selten Cloud-basiert sind. So sind die Unternehmen oftmals angehalten, das Thema KI als Eigenentwicklung mit den vielen Open Source Frameworks voranzutreiben. Problematisch für KMUs ist jedoch, dass sie diese Tools meist nicht nutzen können, denn ein generelles aber auch ein mit der Digitalisierung verbundenes Problem ist der Fachkräftemangel. So werden für die Integration

1 https://www.bvmw.de/themen/mittelstand/zahlen-fakten/

2 https://www-genesis.destatis.de/genesis/online?sequenz=statistikTabellen&selectionname=48121

3 https://www-genesis.destatis.de/genesis/online?sequenz=statistikTabellen&selectionname=48121

4 Gnamm, Kalmbach, Schertler: Von der Vision zur Transformation: Digitalisierung ist Chefsache. Bain & Company, 2018

5 https://www2.deloitte.com/de/de/pages/technology-media-and-telecommunications/articles/ki-studie-2020.html

6 https://www.mittelstand-digital.de/MD/Navigation/DE/Themen/Technologien/Kuenstliche-Intelligenz/kuenstliche-intelligenz.html

7 https://de.statista.com/infografik/16992/umsatz-der-in-deutschland-durch-ki-anwendungen-beeinflusst-wird/

8 https://www.industry-of-things.de/ki-softwaremarkt-umsatz-von-62-milliarden-dollar-zu-erwarten-a-1078580/

9 https://www.handelsblatt.com/technik/forschung-innovation/kuenstliche-intelligenz-warum-kaum-ein-unternehmen-in-deutschland-ki-einsetzt/28735510.html

10 https://www.industry-of-things.de/ki-softwaremarkt-umsatz-von-62-milliarden-dollar-zu-erwarten-a-1078580/

11 https://www.kas.de/documents/252038/7995358/K%C3%BCnstliche+Intelligenz+in+kleinen+und+mittleren+Unternehmen.pdf/1894a732-8ead-46f7-90b4-72c0e1a6fe2b?version=1.1&t=1580810247109

der KI hochqualifizierte Softwareentwickler, Datenanalysten und Machine-Learning-Ingenieure gebraucht, die nur schwer zu finden sind. Ein weiteres Problem ist die oft nicht vorhandene technische Infrastruktur in KMUs. Für verschiedene Anwendungen, wie das Training tiefer neuronaler Netze, wird kostenintensive Hardware wie GPUs benötigt. Aufgrund des Fachkräftemangels, der fehlenden finanziellen Ausstattung und des ungewissen Mehrwerts ist Eigenentwicklung keine Option für KMUs. Wenn es der Datenschutz zulässt, ist eine andere Möglichkeit der Integration von KI in KMUs, auf große Anbieter wie IBM, Microsoft oder Google zuzugehen. Diese bieten Plattformen, zumeist in der Cloud, mit zahlreichen spezialisierten Algorithmen. Im Rahmen von Projekten werden die vielen Funktionen der Plattform so zugeschnitten, dass sie auf die betrieblichen Anwendungsfälle passen. In den meisten Fällen entsteht eine Silo-Lösung für ein einzelnes Problem. Die in Betrieb genommenen Verfahren erfordern eine ständige Wartung, da die Halbwertszeit heutiger KI-Methoden, aufgrund der Dynamik in Märkten und Organisationen, immer kürzer wird. Ein weiterer Aspekt liegt in der Gewährleistung der Aktualität der trainierten Modelle, d.h. einem lebenslangen Lernen der Algorithmen. Dies ist notwendig, damit die Prognoseleistung der Modelle über die Zeit nicht abnimmt. Die Herausforderungen für den Einsatz von KI werden deutlich und sie können nur mit klar definierten **strategischen Unternehmenszielen** [12] angegangen werden.

Mangel an guten Daten: Die meisten Methoden der KI basieren auf statistischen Verfahren über Daten, d.h., mittels maschinellen Lernverfahren werden automatisch Muster in zumeist großen Datenmengen identifiziert und strukturiert. So ist die Verfügbarkeit von ausreichend Daten mit einer hohen Qualität essentiell, aber zumeist nicht gegeben. Dies beginnt mit den generellen, gesetzlichen Gegebenheiten und Bedenken aufgrund des Datenschutzes. So ist ein wesentliches Paradigma der DSGVO die Datensparsamkeit bzw. -minimierung[13], die diametral der Notwendigkeit von Daten für maschinelles Lernen gegenübersteht. Hier besteht Nachbesserungsbedarf seitens der Gesetzgebung, hohen Datenschutz, aber nicht die Datenvermeidung zu adressieren. Unternehmensseitig besteht leider oft wenig Wissen über die vorhandenen Daten, die in den aufgabenzentrisch organisierten Datensilos liegen. Dies können größere Softwaresystem, wie ERP, MES oder CRM-Lösungen, aber auch verteilte Ablagen von Dokumenten oder Spreadsheet-basierte Schatten-IT sein. Die Daten werden selten im Zusammenhang wahrgenommen und so ihr Potential nicht hinreichend evaluiert, entweder aufgrund ihrer Unbekanntheit oder der mangelnden datenzentrischen Kompetenzen bei den Mitarbeitern. So müssen die Unternehmen das strategische Ziel setzen und die nötigen Ressourcen schaffen, einen Datenkatalog zu erstellen, der Daten samt Typ, Zugang und Metriken bzw. Metadaten wie etwa Aktualität, Veränderlichkeit oder potentielle Rechte bzw. Nutzungs-

12 https://www.k-zeitung.de/mittelstand-tut-sich-schwer-mit-kuenstlicher-intelligenz

13 https://www.datenschutzexperte.de/blog/datenschutz-im-unternehmen/datensparsamkeit-und-datenminimierung/

einschränkungen dokumentiert. Neben Herausforderungen bzgl. der Interoperabilität der Daten kann aus solch einem Katalog ein weiteres Datenproblem hervorgehen: KMUs haben oftmals das Problem, allein aufgrund ihrer Größe, dass das Datenpotential für KI per se deutlich kleiner ist als bei einem Großunternehmen, wo auch der digitale Reifegrad fortgeschrittener ist. So müssen bspw. erst die Sensorik an Produktionsstraßen sowie die Hardware für das Sammeln und Speichern geschaffen werden[14]. Eine andere Option, dem Datenmangel Herr zu werden, ist die Verwendung von Daten Dritter. Jedoch hat sich auch hier gezeigt, dass die in (Open) Data Plattformen oder in Partnernetzwerken verfügbaren Daten sich nur bedingt eignen[15]. Die Verfügbarkeit von Daten muss deutlich gesteigert werden, um bessere Wertschöpfungen zu ermöglichen.

Erklärbarkeit und Akzeptanz: Vorteile heutiger KI-Methoden aus dem Bereich des Deep Learnings (DL) sind, dass Muster bzw. Lösungen in Daten gefunden werden, die Experten nur mit sehr hohem Aufwand oder gar nicht finden können. Die Vorhersagen mit DL sind meist präziser als mit herkömmlichen Methoden des maschinellen Lernens. Aber das DL bringt das Problem mit, dass die Methode eine Art Black Box darstellt, die durch den Menschen nicht oder nur mit sehr hohen Aufwänden verstanden und ggf. erklärt werden kann. Gerade diese fehlende Transparenz, wie das Modell zu einer Lösung kommt, ist entscheidend für die Abwägung der Nutzung der KI-Ergebnisse, wie etwa in der Medizin bei der Differentialdiagnose[16]. Fundamental wichtig ist es daher, in bestimmten Anwendungsfällen dem Nutzer eine verständliche Begründung für Entscheidungen aktiv bereitzustellen, um die Ergebnisse nachvollziehen zu können[17].

14 https://www.mittelstand-digital.de/MD/Redaktion/DE/Publikationen/ki-Studie-2021.html

15 https://www.kas.de/de/analysen-und-argumente/detail/-/content/ki-in-kmu-daten-teile

16 https://www.bmwk.de/Redaktion/DE/Artikel/Digitale-Welt/GAIA-X-Use-Cases/differential-diagnose.html

3. Erklärbare KI

Das Forschungsfeld der erklärbaren KI (eXplainable AI, XAI) entwickelte sich stark in den zurückliegenden Jahren mit dem Hype um maschinelles Lernen bzw. Deep Learning, da diese Methoden, im Kontrast zur symbolischen KI, für den Menschen nicht verständlich funktionieren. Doch allein die zugehörige Terminologie für das Forschungsfeld XAI gilt es gut zu differenzieren. Arrieta et al[18]. unterscheidet hierzu treffend:

- **Verstehbarkeit (Understandability):** ein Modell kann seine Funktion einem Mensch vermitteln, ohne die innere Struktur zu erläutern, ohne diese zu erklären.
- **Verständlichkeit (Comprehensibility):** ist die Eigenschaft eines Lernalgorithmus, sein erlerntes Wissen in einer für den Menschen verständlichen Weise darzustellen.
- **Interpretierbarkeit (Interpretability):** ist die Fähigkeit, eine Bedeutung in verständlichen Begriffen für den Mensch zu erklären.
- **Transparenz (Transparency):** ein Modell ist transparent, wenn es unter der Annahme nachvollziehbarer Eingangsdaten selbsterklärend ist.
- **Erklärbarkeit (Explainability):** ist die Bereitstellung einer verständlichen Begründung für das Ergebnis eines Modells für eine Zielperson in einem bestimmten Kontext.

17 https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.html

18 https://arxiv.org/abs/1910.10045

Letztlich führen die Begriffsdefinitionen auf zwei Problemstellungen hin: 1) ist das Modell algorithmisch transparent, daher, es kann nachvollzogen werden, und 2) wenn ein Modell nicht transparent ist, wie kann ich dessen Entscheidungen zielgruppengerecht aufbereitet und verständlich präsentieren? Die unterschiedliche Betrachtung der Zielgruppen ist von enormer Wichtigkeit, da das Hintergrundwissen und die Ziele der Erklärung stark divergieren. So wollen die Nutzer der Modellvorhersagen wissen, ob sie bspw. fair behandelt werden, wohingegen die Entwickler an einer hohen Effizienz und Manager an der Einhaltung von Regularien interessiert sind. Die verschiedenen Ziele der Erklärbarkeit, wie Fairness, Vertrauenswürdigkeit oder das Verständnis zur Robustheit, Übertragbarkeit oder Qualität, werden von Arrieta et al. bzw. den ethischen Richtlinien für vertrauenswürdige KI[19] klar formuliert und abgegrenzt.

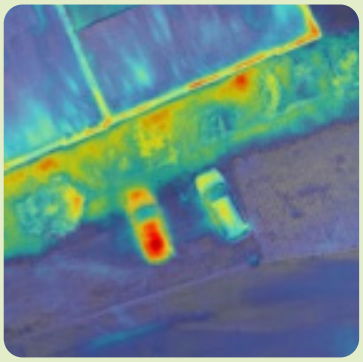
Hieraus ergibt sich die Fragestellung: Wann sollte ich welche Strategie bzw. Technologie einsetzen, um eine Erklärbarkeit für die KI zu gewährleisten? Hierzu kann auf die Orientierungshilfe seitens der BMWi-Studie "Erklärbare KI" verwiesen werden, bei der zwischen den wesentlichen Zielgruppen der KI-Entwickler sowie der KI-Nutzer/Domänenexperten unterschieden wird. Arbeitet bspw. letztere Gruppe mit Bilddaten, so haben sich Methoden aus dem Bereich der **Saliency Maps**[20], wie z.B. GradCam[21], etabliert. Salienz bezieht sich auf einzigartige Merkmale (Pixel, Auflösung usw.) des Bildes im Zusammenhang mit der visuellen Verarbeitung. Diese einzigartigen Merkmale stellen die visuell anziehenden Stellen in einem Bild dar. Es werden über den zu untersuchenden Bildern sogenannte Heatmaps angezeigt, die für das Modell interes-

19 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

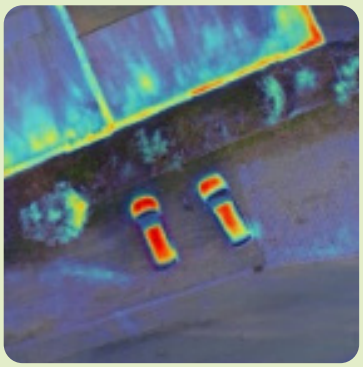
20 https://analyticsindiamag.com/what-are-saliency-maps-in-deep-learning/

21 Selvaraju et al: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision (IJCV). 2019. https://arxiv.org/abs/1610.02391

Eingangsbild



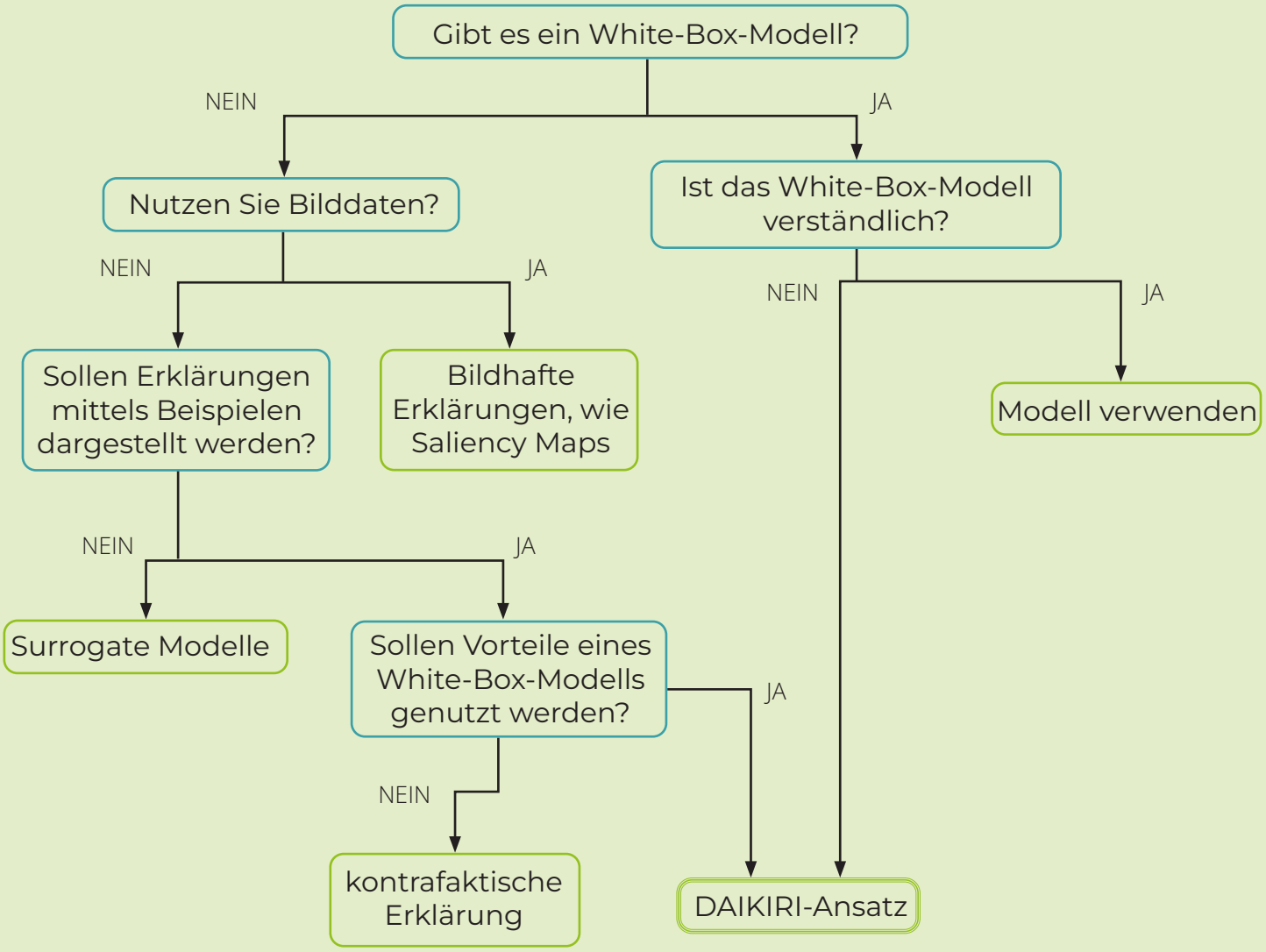
Saliency Map -I_{rgb} als Eingabe



Saliency Map -I_{max} als Eingabe

Beispiel für Saliency Maps für den Anwendungsfall thermale Anomalieerkennung.

Orientierungshilfe für Strategien und Werkzeuge von erklärbarer KI für Domainenexperten



sante Regionen hervorheben (siehe Grafik). Werden die Erklärungen auf strukturierten Daten benötigt, so kommen andere Methoden in Frage. Sogenannte Stellvertretermodelle (Surrogat Models[22]) sind stark vereinfachte, parallel trainierte Modelle, die zwar nur eine Näherungslösung abbilden, aber deren Entscheidungslogik deutlich leichter für den Domänenexperten verständlich sind.

4. Der DAIKIRI-Ansatz im Überblick

Trotz der Entwicklung im Bereich der Erklärungswerkzeuge für Black-Box-Modelle, existieren bisher nur wenig gute Werkzeuge, die intuitiv verständliche Entscheidungserklärungen liefern. Insbesondere die bei einigen Anwendungsfällen wichtigste Zielgruppe der Erklärungen, die Domänenexperten, sind davon betroffen. Laut einer Studie des BMWi zu Erklärbarer KI sollten bei Anwendungen, die hohe Anforderungen an die Nachvollziehbarkeit der Modellentscheidungen stellen, bevorzugt White-Box-Modelle verwendet werden, wenn diese im Vergleich zu Black-Box-Ansätzen ausreichend gute Ergebnisse liefern[23]. Deshalb ist es erstrebenswert, White-Box-Modelle zu verbessern, weiterzuentwickeln und existierende Verfahren auf weitere Datentypen zu überführen.

White-Box-Modelle sind in Bezug auf ihre Logik und ihre Wirkungsmechanismen für Menschen direkt nachvollziehbar. Speziell werden beim strukturierten maschinellem Lernen, kurz SML, aus Wissensgraphen intrinsisch erklärbare Modelle gelernt, die zudem verbalisiert werden können, das bedeutet, in natürliche Sprache überführt werden können. Dieser Ansatz ist die Idee hinter dem gemeinsamen Forschungsprojekt DAIKIRI[24] der DICE Gruppe der Uni Paderborn, der USU Software AG, der elevait GmbH & Co. KG sowie der pmOne AG. DAIKIRI steht dabei für Erklärbare Diagnositische KI für industrielle Daten.

Um SML anwenden zu können, müssen die Daten als Wissensgraph vorliegen. Die Schwierigkeit hierbei ist, dass Industrie- und Anlagendaten meist eine sehr flache Datenstruktur aufweisen. Die flachen Daten müssen zunächst annotiert werden, sodass sie als Wissensgraph repräsentiert werden können. Dieses Überführen der Daten erfolgt bisher zumeist manuell. Zudem sind für diesen sogenannten Semantifizierungsprozess Domänenexperten nötig, welche aber von Haus aus keine Logikexperten sind, wie es für die Generierung logischer Ausdrücke (Axiome) von Nöten wäre.

Daher hat sich das Projekt DAIKIRI zum Ziel gemacht, neue, effiziente Verfahren zur semi-automatischen Semantifizierung von flachen Daten mittels aktiven Lernens und SML zu entwickeln und automatisch natürlichsprachige Erklärungen für die Klassifikationsergebnisse von SML sowie Black-Box Algorithmen zu generieren.

Um dieses Vorhaben umzusetzen, wurde die Open Source Bibliothek Vectograph zur automatischen Generierung von Graphdaten aus gegebenen tabellarischen Daten entwickelt[25]. Außerdem wurden neue, effiziente Embeddingverfahren für Wissensgraphen entwickelt, die auf große Datenmengen skalieren, um die Daten einzubetten[26] [27]. Ein neuartiges Clusteringverfahren gruppiert

22 Alizadeh, R., Allen, J.K. & Mistree, F. Managing computational complexity using surrogate models: a critical review. Res Eng Design 31, 275–298 (2020). https://doi.org/10.1007/s00163-020-00336-7

23 https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.html

24 https://daikiri-projekt.de/

25 https://github.com/dice-group/vectograph

26 C. Demir and A. N. Ngomo. Convolutional complex knowledge graph embeddings. In ESWC, volume 12731 of Lecture Notes in Computer Science, pages 409–424. Springer, 2021.

27 C. Demir, J. Lienen, and A. N. Ngomo. Kronecker decomposition for knowledge graph embeddings. In HT, pages 1–10. ACM, 2022

Datenpunkte mit ähnlichen Charakteristiken basierend auf ihren Embeddings[28]. Die Benennung der Cluster erfolgt semi-automatisch. Dabei wird der häufigste Typ an alle Datenpunkte eines Clusters weitergegeben. Im Nachgang kann diese Benennung durch einen Human-in-the-Loop Ansatz manuell modifiziert werden und, wenn gewünscht, manuell Querverbindungen hergestellt werden[29]. Über den anschließenden Axiom Generator wird schließlich die Ontologie erstellt und in einer Datei abgespeichert. Um auf dieser Ontologie ein SML Modell über die Ontolearn Bibliothek[30] zu lernen, müssen zusätzlich positive und negative Beispiele für die Klassifikation vorliegen. Liegen für einen Use Case keine vollständig gelabelten Daten vor, kann an dieser Stelle der hybride Ansatz gewählt werden, dazu wird ein Unsupervised Black-Box-Modell gelernt. Die Klassifikationsergebnisse hiervon werden dann gemeinsam mit der Ontologie verwendet, um ein SML Modell zu lernen. Die gelernten Konzepte können anschließend mittels des DAIKIRI Verbalizers[31] in natürliche Sprache überführt werden.

Im Rahmen des Forschungsprojekts wurden die einzelnen Komponenten in die DAIKIRI-Plattform integriert und zu einer Pipeline zusammengeführt. Diese Pipeline wurde auch für einen realen Anwendungsfall zur Logistik von kleinen Teilen in einer Fertigungshalle umgesetzt. In diesem Use Case geht es darum anormale Änderungen der Füllstände in den Lagerboxen zu erkennen und zwischen drei verschiedene Anomalietypen zu unterscheiden.

Es hat sich gezeigt, dass die neu entwickelten Methoden auch für reale Daten geeignet sind und auf große Datenmengen skalieren. Unter Verwendung von zusätzlichen Querverbindungen konnten gute Konzepte über den White-Box-Ansatz generiert werden. Hier gilt es in Zukunft Methoden für eine automatisierte Erstellung von Querverbindungen zu entwickeln. Die von SML gelernten Konzepte sind über die Verbalisierung für Menschen, insbesondere Domänenexperten gut verständlich. In ihrer Art sind sie aber nicht vergleichbar mit Entscheidungserklärungen von Black-Box-Modellen. Als Vorteil von White-Box-Modellen sei nochmal ihre intrinsische Erklärbarkeit hervorgehoben. Modell und Erklärungen passen damit immer zusammen. Dadurch entfällt die Anforderung, dass neben der Güte der Modellvorhersagen auch die Zuverlässigkeit der Erklärungen gegeben sein muss, wie dies beim Black-Box-Ansatz der Fall ist.

28 H. M. Zahera, S. Heindorf, S. Balke, J. Haupt, M. Voigt, C. Walter, F. Witter, and A.-C. Ngonga Ngomo. Tab2onto: Unsupervised semantification with knowledge graph embeddings. In European Semantic Web Conference, pages 47–51. Springer, 2022.

29 H. M. Zahera, S. Heindorf, and A.-C. N. Ngomo. Asset: A semi-supervised approach for entity typing in knowledge graphs. In Proceedings of the 11th on Knowledge Capture Conference, pages 261–264, 2021.

30 S. Heindorf, L. Blübaum, N. Düsterhus, T. Werner, V. N. Golani, C. Demir, and A. N. Ngomo. Evolearner: Learning description logics with evolutionary algorithms. In WWW, pages 818–828. ACM, 2022.

31 https://github.com/dice-group/daikiri-verbalizer