Flexibilisierung von Assistenzsystemen mittels Machine Learning in der Steuerung technischer Systeme

Jörg R. RUDNICK, Benjamin WEYERS

Human-Computer Interaction, Universität Trier
Behringerstraße 21, D-54296 Trier

Kurzfassung: Wir suchen ein Assistenzsystem, welches nicht rezeptgebundene Fertigkeiten von Operateuren im Sinne einer nicht angetasteten Handlungskomponente einbindet. An Stelle von Standard Operation Procedures als Entscheidungsgrundlage für die Assistenz verwendet unser Ansatz Reinforcement Learning, insbesondere auch dessen Modellbildung nutzend, um die Grenze maschinellen Eingreifens auszumachen.

Schlüsselwörter: Assistenzsysteme, Reinforcement Learning, SOP

1. Einführung

Aus der Warte der Gestaltung von Assistenzsystemen für die Steuerung technischer Systeme haben sich **Standard Operating Procedures** (SOP) als mögliche Operationalisierungsgrundlage erwiesen (Weyers et al. 2015; Weyers et al. 2020) - doch können diese nur eine Teilmenge des Aufgabenfeldes erfassen: Insbesondere für kritische Umgebungen, wo die Restabilisierung von Zuständen eines zu steuernden technischen Systems außerhalb eines planmäßigen Bereiches gefragt ist, wären Assistenzsysteme von Interesse. Gerade solche Zustände lassen sich mit rezeptgeführtem Vorgehen schwer vereinbaren. So stellt sich die Frage nach der Möglichkeit von Unterstützung für nachhaltig menschliche und weniger automatisierbare Fertigkeiten, wie z.B. Improvisation oder Kommunikation für solche unscharfe und kritische Ausnahmesituationen.

Als glaubwürdiger Ansatz erscheint uns hierzu, einen entsprechenden Kern-bereich nicht zu automatisierender Fertigkeiten wirklich auszuschließen, und das Assistenzsystem vielmehr auf einen verbleibenden Rest Einfluss nehmen zu lassen. Unter diesen Bedingungen würde unser Assistenzsystem also intuitivere und kreativere Arbeiten unterstützen können, indem die Einflussnahme sich nur mit der Ausräumung von mentalen Hindernissen beschränkterer Komplexität befasst.

Ziel dieses Artikels ist es ein Konzept für ein solches Assistenzsystem auf Basis von Reinforcement Learning (RL) vorzustellen. Der Einsatz von RL in Assistenzsystemen ist schon zuvor untersucht worden - jedoch ohne Bezug auf Operator-Training, und eher im Sinne einer rein technischen KI-Unterstützung, und nicht als Modell-Grundlage, wie in dieser Arbeit (Neelakantam et al. 2020; Chen & Soh 2017).

2. Grundlagen und Reinforcement Learning

Im Rahmen dieser Arbeit bezeichnen wir die Menge der Zustände des zu bedienenden Systems mit $s \in S$, die durch den Operator ausführbaren Handlungen mit $a \in A$, und das daraus resultierende Systemverhalten $\Pi := S^A$, als Menge von

Abbildungen von S auf A , wobei wir ein Element $\pi \in \Pi$ als **Policy** bezeichnen. Als Entscheidungsgrundlage für die Gestaltung des schlussendlichen Algorithmus legen wir einen Ertrag $r \in R$ für jeden Zeitschritt t zugrunde, welcher von S abhängig und bereits im Vorfeld vereinbart sei.

Der Kern unseres Ansatzes hat zum Ziel, RL hinsichtlich einer Ergänzung zu bestehenden Assistenzsystem-Ansätzen zu untersuchen. Dabei wird eine Abgrenzung der maschinell zu betreuenden Handlungen erst denkbar durch dessen weitgehende **Freiheit von π-Vorwegannahmen**. Hierzu werden Policies unmittelbar aus Handlungen in Abhängigkeit von Bewertungen und Dynamik der System-zustände approximativ gebildet. Dies kann grob anhand der folgenden 2-Schritt-Iteration illustriert werden: Die Bewertung eines Zustandes lässt sich verbessern als gewichtetes Mittel über die empfohlenen Handlungen a der gewichteten Mittel über erwartbare Erträge r und Folgezustände s' solcher Handlungen der diskontierten ursprünglichen Bewertung dieser Folgezustände zuzüglich des Schritt-Ertrages, indem die Schätzung teilweise durch exakte Werte ersetzt ist:

$$v'(s) := \sum_{s'} \sum_{r} P\left(s', r | s, \pi(s)\right)\left(r + \gamma v(s')\right) .$$

Mit diesem rekursiv erfassten $v'(s)$ kann mit dem Policy-Improvement-Theorem nach Bellman (1957) und Howard (1960) eine verbesserte Policy als Handlung mit dem höchsten gewichteten Mittel über erwartbare Erträge r und Folgezustände s' solcher Handlungen der diskontierten ursprünglichen Bewertung dieser Folgezustände zuzüglich des Schritt-Ertrages abgeleitet werden, wobei γ eine Rate der Diskontierung von Ertrag über die Zeit ist, um einen Ertrags-Begriff über alle Zeiten ableiten zu können:

$$\pi'(s) := argmax_a \sum_{s'} \sum_{r} P\,(s', r | s, a)\left(r + \gamma v'(s')\right).$$

Auf solche Weise gelingt es dem RL, das Kontinuum kognitiver Leistungen weitestgehend willkürfrei zu modellieren. Man kann sich dazu das RL als systematische Erschließung des Suchraums aller Algorithmen lediglich anhand der Ertrags-Metrik für einen Schritt vorstellen. Dieses iterativ-approximative Vorgehen ist ebenfalls interessant durch die Aussicht, lediglich über die Schrittzahl eine Art Ordnungsrelation ableiten zu können.

Schließlich bietet eine probabilistische Fundierung im RL die Aussicht, dass gerade in Hinblick auf fortgeschrittene Verfahren für die effiziente Durchquerung des Suchraums leistungsfähige Ansätze vorliegen, welche für Trade-offs gezielt Leistungs-/Aufwands-Verhältnisse abwägen, insbesondere unter Anwendung des Policy-Gradient-Theorems (Marbach & Tsitsiklis 1998) - was in Verbindung mit einer wie oben hergeleiteten Ordnungsrelation dann genau zu einer brauchbaren Metrik für ‚allgemeine kognitive Leistung' führen könnte, welche für unser Vorhaben von grundlegender Bedeutung und Gegenstand aktueller Untersuchung ist.

3. Bezugsfeld

Für die Darlegung des Ansatzes erwies es sich als kritisch, mit größtmöglicher Konsequenz jegliche Szenarien auszuschließen, welche die Komplexität der Modellierung unnötig ausufern lassen - und gleichzeitig sicherzustellen, dass ein nicht triviales Beispiel noch vorliegt. Beides betrachten wir als Vorerwägungen.

3.1 Hinreichend komplexes technisches Arbeitsproblem

Aufgrund der steigenden Leistungsfähigkeit von Computerssystemen wird es wichtig, den Begriff ‚menschliche und weniger automatisierbare Fertigkeiten' weiter zu operationalisieren. Wir gehen von der Arbeitshypothese aus, dass die Komplexität chaotischer bzw. seltsamer Attraktoren diesem weitgehend genügt (z.B. für logistische Gleichung González & Pino 1999), indem so die Erfordernis erhöhter Flexibilität typisch menschlicher Einsatzverhältnisse gerechtfertigt werden kann. Uns erscheint eine sehr einfache 2-zyklische Bifurkation, d.h. ein Pendeln zwischen zwei Zuständen, unter bestimmten Umständen noch als brauchbarer Kompromiss, nämlich dann, wenn die zugrundeliegende Dynamik über ein Parameter vergleichbar einer Reynolds-Zahl bestimmt ist, dessen Änderung einen gleitenden Übergang zu Zyklen beliebig vieler Zustände bzw. seltsamen oder chaotischen Verhaltens erlaubt.

Konkret beziehen wir uns auf die entsprechende Anpassung eines einfachen Modells eines Kernkraftwerkes (Weyers et al. 2017) - bei welchem sich dann eine 2-zyklische Bifurkation als ein oszillierender Betrieb zwischen zwei Niveaus darstellen lässt, z.B. auf durchaus nutzbringende Weise wie bei einem TRIGA-Reaktor (Dyson 1981). Der Zweckbestimmung unserer Untersuchung gemäß wählen wir eine RL-Lösung, welche noch schwach genug ist, um solche Oszillation nicht zu erkennen - und somit nicht abwechselnden Pfaden folgt. Dies kann sich dann, je nach Reward-Vereinbarung für das RL, weniger günstig auswirken beispielsweise durch Aufwand unbegründeter Versuche, den Prozess in eine einheitliche Bahn zu zwingen, oder Fokusverlust bei einem Verbleib im Schwankungsbereich beider Phasen.

3.2 Lerndaten von Probeläufen mit Menschen

Für Erweiterte Assistenzsysteme, welche die menschliche Leistung nicht vorwegnehmen, ist die Verfügbarkeit ausreichender Lerndaten mit Probeläufen mensch-licher Teilnehmer eine gegebene Voraussetzung.

Für die Vergleichbarkeit ist zudem wichtig, dass die Zielsetzungen der menschlichen Spieler und des RL-Systems einander weitgehend entsprechen. Hierzu haben wir zu Beginn der menschlichen Versuchsreihen einerseits die **Rewards** - d.h., die Bewertungen der Ereignisse in derselben Zeiteinheit - als auch γ, die Discountrate welche das Gewicht der Rewards mit jeder Zeiteinheit abschwächt, und so die Ableitung eines über die gesamte Zukunft gehenden Wertes v (**Value**) erlaubt.

Prämisse, Lern-Daten: Wir setzen voraus, dass ein hinreichender Bestand an Testdurchläufen menschlicher Spieler vorliegt, und für die **value**-Ermittlung ein geeignetes γ vereinbart ist.

Der tatsächliche Bedarf an Teilnehmerdaten ist in dem frühen Stadium unserer Studie noch nicht klar abzuschätzen, und wird wie auch die Findung geeigneter Discountraten Gegenstand für Folgeuntersuchungen sein.

4. Ansatz

In diesem Abschnitt wird ein grundlegender Ansatz eines Erweiterten Assistenzsystems präsentiert, der nicht mehr trivial und entwicklungsfähig ist. Weiterhin zeigen wir, dass ein solches Assistenzsystem im Wesentlichen der Anwendung eines RL entspricht.

4.1 Übersicht

Wir fassen ein Erweitertes Assistenzsystem als eine Abbildung auf, die bei gegebenem Zustand des zu steuernden Systems und gegebener Handlung eines menschlichen Nutzers gegebenenfalls eine Hilfestellung erwidert,

$$assistenz: S \times A \rightarrow Maybe\ Hilfestellung.$$

Die Hilfestellung würde naturgemäß aus einer Abbildung auf dem Systemzustand, sowie der angestrebten Handlungen von menschlichem Teilnehmer sowie des Assistenzsystems hervorgehen,

$$hilfe: S \times A \times A \rightarrow Hilfestellung.$$

Da die schlussendliche Präsentation einer solchen Hilfestellung nicht Thema dieser Untersuchung ist, wollen wir uns o.B.d.A. mit der einfachsten Variante begnügen, d.h. der rohen Übergabe der Handlungsempfehlung des Assistenz-systems an den Nutzer,

$$Hilfestellung := A.$$

Wir wollen weiterhin voraussetzen, dass die möglichen Handlungen einem metrischen Raum angehören, so dass auf ihnen eine bereits bekannte Abstands-funktion definiert ist, welche ausdrückt, wie erheblich der Unterschied zwischen zwei Handlungen in Bezug auf die Folgen ist,

$$abstand: A \times A \rightarrow \mathbb{R}_+.$$

Tatsächlich könnte man auch versuchen, diese Abstandsfunktion ebenfalls über das verwendete RL abzuleiten, was interessanter Gegenstand einer Folgestudie wäre.

Die Veranlassung einer Hilfestellung soll nun im Folgenden in Abhängigkeit von Zustand und Abweichung der gewählten Handlung des menschlichen Teilnehmers von jener des Assistenzsystems gesehen werden. Das bedeutet, dass nur in einen Zustand, für welchen sich die Stärke des Assistenzsystems gezeigt hat, und für die eine ausreichende Handlungsabweichung identifizierbar ist, einen Eingriff rechtfertigt. Zusätzlich wollen wir feststellen, dass für diese Funktion Zustand und Abstand als unabhängig voneinander angesehen werden können, so dass wir stattdessen eine AND-Verknüpfung zweier Funktionen verwenden können,

$$hilfebedarfAbst: \mathbb{R}_+ \rightarrow Bool,$$
$$hilfebedarfZust: S \rightarrow Bool.$$

Hier kann **hilfebedarfAbst** wiederum als einfacher Schwellwert-Vergleich ange-sehen werden, so dass zuletzt noch eine Funktion verbleibt, welche in der Lage ist, zu einem Zustand eine Handlungsempfehlung abzugeben, was im RL natürlich nichts anderes als eine **Policy** $\pi: S \rightarrow A$ ist.

Wie man feststellt, kann das Assistenzsystem als ein Wrapper um die beiden wirklich erheblichen Funktionen **policy** und **hilfebedarfZust** aufgefasst werden, welche nunmehr im Wesentlichen die vormals von SOPs übernommenen Funktionen beinhalten – ihnen werden wir uns im Folgenden widmen.

4.2 hilfebedarfZust

Zu Beginn soll darauf eingegangen werden, wieso in dieser Arbeit noch nicht auf eine Hilfebedarfsentscheidung unter Einbezug der Nutzerhandlung eingegangen wird, obwohl es doch Sinn macht, davon auszugehen, dass unter bestimmten Umständen bestimmte Handlungen offenkundiger als anderer Unterstützungsbedarf anzeigen, z.B. wenn eine Handlung sich für einen Zustand als offenkundig absurd darstellt.

Ein Ansatz wäre, sich an einer Funktion

$$hilfebedarf: S \times A \rightarrow Bool$$

zu versuchen. Das dabei entstehende Problem ist, dass Voraussetzung wäre, entweder (a) mittels Machine Learning aus begrenzten Teilnehmerdaten eine lückenlose Klassifikation auf ganz $S \times A$ ableiten zu können, oder (b) den **Value** menschlicher Teilnehmerhandlung zur Laufzeit des Assistenzsystems zu ermitteln. Letzteres machte streng genommen in dem vereinbarten Rahmen nur Sinn, wenn man eine Ermittlung unter Umgehung von Handlungsempfehlungen selbst deutlich effizient erbrächte, wofür uns kein Verfahren bekannt ist. Aber auch ersteres dürfte noch unseren gegenwärtigen Rahmen sprengende Arbeiten erfordern.

Eine ähnliche Herangehensweise wäre eine zustandsbedingte Abstandsfunktion zu finden, die je nach Lage unterschiedliche Sensibilität für Handlungsabweichungen erlaubt. Dies könnte ein weniger aufwändiger Ansatz sein, da die Value-Berechnung im RL bereits tief verankert ist, ginge aber trotzdem über den Rahmen dieser einführenden Arbeit hinaus.

Wir begnügen uns hier mit einer Partitionierung bezüglich Hilfestellungsbedarfs auf Grundlage allein der Anforderungssituation. Das Vorgehen ergibt sich dann wie folgt:

1. Durchgehende Ermittlung der Values aus Rewards auf allen menschlichen Teilnehmerdaten, durch Diskontierung mittels der Rate γ.
2. Stückelung der Trajektorien in Augenblicks-Schnappschüsse aus Grundlage des Zustandes mit Value.
3. Ermittlung der jeweiligen Handlungsempfehlung des Assistenzssystems durch Aufrufen von **policy**.
4. Ermittlung der Values für diese Empfehlungen mittels der **policy** zugeordneten **actionValue$_{\text{policy}}$**-Funktion, $actionValue_{policy}: S \times A \rightarrow Value$.
5. Ausrechnen der Differenzen zwischen beiden und Diskretisierung dieser zu Boole'schem Wahr/Falsch.
6. Ggf. noch Glättung dieser Partitionierung mit Hilfe weitere Machine Learning-Verfahren.

4.3 policy & actionValue$_{policy}$ – das Reinforcement Learning

Mit **policy**- & **actionValue$_{\text{policy}}$** als einzigen verbleibenden nicht trivialen Funktionen haben wir das Erweiterte Assistenzsystem auf eine RL-Aufgabe reduziert, da man dies grob als Sammlung von Verfahren auffassen kann, geeignete **policy**- & **actionValue$_{\text{policy}}$**-Ausprägungen im Wechselspiel iterativ zu ermitteln.

Zwei weitere Eigenschaften sind nun noch festzustellen, welche einen Einsatz für Erweiterte Assistenzsysteme untermauern. Diese wollen wir **vertikale Variierbarkeit** und **horizontale Homogenität** nennen.

Bemerkung, Vertikale Variierbarkeit: Das zu erwartende Value-Niveau von RL-Systemen kann als quasi-kontinuierlich einstellbar angesehen werden.

Begründung: Dank des iterativen Ansatzes des RL kann der zu erwartende Value beispielsweise über die Wahl der Anzahl der Lernschritte eingestellt werden.

Bemerkung, Horizontale Homogenität: Homogenität im Sinne der Vermeidung einer nicht leistungsbezogenen Bevorzugung von Policies wird durch den Einsatz probabilistischer Verfahren begünstigt.

Begründung: Für probabilistische Verfahren ist es üblich, dass der Suchraum ausgewogen durchlaufen wird (vgl. etwa Markov Chain Monte Carlo (Gamerman & Freitas 2006)); es wird weitestgehend auf Anfangs-Hypothesen & -Heuristiken verzichtet, welche die Gefahr eines Bias beinhalten könnten.

Abschließend bleibt noch zu besprechen, was nun an Stelle der vormaligen Standard Operation Procedures als Steuerparameter in ein Erweitertes Assistenzsysteme einzugehen hat. Neben den schon anfangs erwähnten Teilnehmerdaten werden dies genau jene Daten sein, welche auch einem RL eine Richtung vorgeben - im Wesentlichen also eine **Reward**-Vereinbarung und eine Discountrate γ.

5. Schlussfolgerung und Implementation

Motiviert durch die beschriebenen Einsichten haben wir mit empirischen Arbeiten begonnen, um herauszufinden, inwieweit Erweiterte Assistenzsysteme in dem beschriebenen Sinne in der Lage sind, sich in der Praxis zu bewähren: Anhand eines Prototypen und auf Grundlage von verfügbaren menschlichen Operatordaten wird anhand begrenzt komplexer Sachverhalte einerseits die Separierbarkeit menschlicher kognitiver Leistung bezüglich einer Unterstützbarkeit, wie auch die grundlegende Brauchbarkeit eines darauf aufbauenden Assistenzsystems untersucht. Dazu untersuchen wir einen Policy Gradient Method-Ansatz, um die Möglichkeiten der Bildung einer Abgrenzungsmetrik möglichst tiefen Einblick zu gewinnen.

6. Literatur

Bellman RE (1957) Dynamic Programming. Princeton University Press 1957.

Chen H, Soh YC (2017) A Cooking Assistance System for Patients with Alzheimers Disease Using Reinforcement Learning. IntJ Inf Tech Vol. 23, No. 2 2017.

Dyson FJ (1981) Disturbing the Universe. Basic Books, Sloan Foundation Science Series, 1981.

Gamerman D, Freitas H (2006) Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Texts in statistical science 68, Chapman & Hall/CRC 2006.

González JA, Pino R (1999) A random number generator based on unpredictable chaotic functions, Computer Physics Communications, S. 109-114, Vol. 120 2-3, Aug 1999, Elsevier.

Howard R (1960) Dynamic Programming and Markov Processes, MIT Press, Cambridge MA.

Marbach P, Tsitsiklis JN (1998) Simulation-based optimization of Markov reward processes. MIT Technical Report LIDS-P-2411.

Neelakantam G, Onthoni DD, Sahoo PK (2020) Reinforcement Learning Based Passengers Assistance System for Crowded Public Transportation in Fog Enabled Smart City. Electronics 2020, 9, 1501.

Weyers B, Frank B, Bischof K, Kluge A (2015) Gaze Guiding as Support for the Control of Technical Systems. IntJ Inf of Systems f Crisis Response and Mgmt, 7(2), S. 59-80, April-June 2015.

Weyers B, Harrison MD, Bowen J, Dix A, Palanque P (2017) Case Study 1–Control of a Nuclear Power Plant. In: Weyers B, Bowen J, Dix A, Palanque P (Eds.), The Handbook of Formal Methods in Human-Computer Interaction, 4.2, S. 91-101, Springer 2017.

Weyers B, Frank B, Kluge A (2020) A Formal Modeling Framework for the Implementation of Gaze Guiding as an Adaptive Computer-Based Job Aid for the Control of Complex Technical Systems, IntJ of Human-Computer Interaction, 36:8, S. 748-776 2020.

Danksagung: Diese Arbeit wurde unterstützt von der DFG (Deutsche Forschungsgemeinschaft. Fördernummer WE5408/3-1).

Gesellschaft für
Arbeitswissenschaft e.V.

Arbeit HUMAINE gestalten

67. Kongress der
Gesellschaft für Arbeitswissenschaft

Lehrstuhl Wirtschaftspsychologie (WiPs)
Ruhr-Universität Bochum

Institut für Arbeitswissenschaft (IAW)
Ruhr-Universität Bochum

3. - 5. März 2021

GfA-Press

Bericht zum 67. Arbeitswissenschaftlichen Kongress vom 3. - 5. März 2021

Lehrstuhl Wirtschaftspsychologie, Ruhr-Universität Bochum
Institut für Arbeitswissenschaft, Ruhr-Universität Bochum

Herausgegeben von der Gesellschaft für Arbeitswissenschaft e.V.
Dortmund: GfA-Press, 2021
ISBN 978-3-936804-29-4

NE: Gesellschaft für Arbeitswissenschaft: Jahresdokumentation

Als Manuskript zusammengestellt. Diese Jahresdokumentation ist nur in der Geschäftsstelle erhältlich.

Schriftleitung: Matthias Jäger
im Auftrag der Gesellschaft für Arbeitswissenschaft e.V.

Screen design und Umsetzung
© 2021 fröse multimedia, Frank Fröse
office@internetkundenservice.de · www.internetkundenservice.de