Karlsruher Institut für Technologie

Institut für Theoretische Informatik (ITI)
Anwendungsorientierte Formale Verifikation
Prof. Dr. Bernhard Beckert

Maschinelles Prüfen von Fairness-Eigenschaften mit Hilfe formaler Informationsflussanalyse

1 Hintergrund

Immer häufiger werden relevante Entscheidungen über Resourcenverteilungen nicht mehr von Menschen, sondern von Algorithmen getroffen. Diese ziehen dabei eine Menge von numerischen Merkmalen zu Rate, um eine Entscheidungsempfehlung zu treffen. Solche Algorithmen werden bereits heute in Bereichen wie der Vergabe von Krediten, Risikobeurteilung von Straftätern und weiteren kritischen Aufgabenfeldern eingesetzt.[1]
Hierbei werden häufig viele verschiedene Merkmale berücksichtigt. Allerdings sollen Resourcen auch fair verteilt werden, insbesondere darf dabei nicht unzulässig diskrimiert werden (siehe z. B. COMPAS[2]). Ein mögliches Fairnesskriterium wäre beispielsweise, dass eine Entscheidung nicht von geschützten Merkmalen wie z. B. Hautfarbe abhängen darf.

2 Aufgabe

Ziel dieses Projektes ist es, programmierte Entscheidungsalgorithmen durch maschinelle Informationsflussanalysewerkzeuge auf ausgewählte Fairness-Eigenschaften zu prüfen.
Ihre Aufgabe besteht darin, am Lehrstuhl entwickelte Verifikationstechniken [1, 2] zur präzisen (statischen) Verifikation von Informationsflusseigenschaften auf Fairnesseigenschaften anzuwenden. Neben der Anpassung bestehender Techniken (bspw. KeY) umfasst dies insbesondere auch die Untersuchung wie Fairnesseigenschaften geeignet formalisiert werden können.

3 Abgrenzung

In diesem Projekt wollen wir uns darauf beschränken, programmierte Entscheidungsverfahren zu analysieren. Entscheidungsalgorithmen, die mit Hilfe von Maschinellem Lernen erzeugt werden, sollen für den Moment ausgeklammert werden. Stattdessen soll mit einfachen Beispielen begonnen werden, die wir zu gegebenem Zeitpunkt erweitern werden.

4 Kontakt / Betreuung

Michael Kirsten, kirsten@kit.edu, Raum 228 (Geb. 50.34)
Jonas Klamroth, klamroth@fzi.de, Raum 1.1.27 (FZI)
Dr. Mattias Ulbrich, ulbrich@kit.edu, Raum 229 (Geb. 50.34)

Literatur

[1] Bernhard Beckert u. a. „Using Theorem Provers to Increase the Precision of Dependence Analysis for Information Flow Control“. In: *20th International Conference on Formal Engineering Methods*. LNCS. Springer, 2018. DOI: 10.1007/978-3-030-02450-5_17.

[2] Christoph Scheben und Simon Greiner. „Information Flow Analysis“. In: *Deductive Software Verification - The KeY Book: From Theory to Practice*. Bd. 10001. LNCS. Springer, 2016. Kap. 13. DOI: 10.1007/978-3-319-49812-6_13.

[1] https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses
[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing