WIRTSCHAFT
HOCHSCHULE MAINZ
UNIVERSITY OF
APPLIED SCIENCES

Exposé zur Bachelorarbeit

Studiengang Wirtschaftsinformatik B.Sc. dual

Evaluation der Reinforcement Learning Algorithmen SARSA und Q-Learning am Beispiel eines Strategiespiels

Hochschule Mainz

University of Applied Sciences

Fachbereich Wirtschaft

Vorgelegt von:	Jonas Bingel
	[redacted]
	[redacted] Wiesbaden
	Matrikel-Nr. [redacted]
Vorgelegt bei:	Prof. Dr. Frank Mehler
Eingereicht am:	07.12.2021

Beschreibung des Themas

Reinforcement Learning (RL) ist ein Teilgebiet des Machine Learning, das sich mit der Erstellung von Agenten beschäftigt, die Entscheidungen treffen, um den Gesamtgewinn zu maximieren. Aufgrund des sequentiellen Aufbaus und der Reproduzierbarkeit von (Strategie-)Spielen, bieten diese eine gute Anwendung zur Auswertung von RL Algorithmen. In der Bachelorarbeit sollen die RL Algorithmen SARSA und Q-Learning verglichen und deren Unterschiede verdeutlicht werden. Ferner soll deren Spielstärke, Spielverhalten sowie Performance am Beispiel des simplen Strategiespiels Tic-Tac-Toe evaluiert werden.

Forschungsfrage

Das Ziel der Arbeit ist eine Erarbeitung der Algorithmen SARSA und Q-Learning sowie deren Unterschiede. Am Beispiel des simplen Strategiespiels Tic-Tac-Toe sollen Spielstärke, Spielverhalten und Performance von SARSA und Q-Learning evaluiert werden. Im Zuge dessen sollen die folgenden weiteren Fragestellungen beantwortet werden:

- Wie viele Trainingsepisoden sind jeweils notwendig, um das Spiel durch Self-Play zu erlernen?
- Welche Effekte haben Veränderungen der Hyperparameter, insbesondere der Reward, auf das Training und die Stärke bzw. das Spielverhalten der Agenten?
- Welche Metriken können zur Bewertung der Performance genutzt werden?

Methodik und Vorgehen

Zur Beantwortung der Fragestellung werden die Algorithmen SARSA und Q-Learning erläutert und deren Unterschiede anhand von Beispielen verdeutlicht. Anschließend sollen beide Algorithmen in Java implementiert werden mit dem Ziel eine ideale Spielstrategie für Tic-Tac-Toe zu ermitteln. Die Agenten sollen das Spiel durch Self-Play erlernen, d. h. der Agent spielt nicht gegen andere Agenten, sondern nur sich selbst und bekommt kein externes Know-How bereitgestellt. Zur Evaluation der trainierten Agenten werden Testspiele gegen einen perfekt spielenden MiniMax-Agenten durchgeführt. Für jeden Algorithmus sollen verschiedene Kombinationen von Hyperparametern getestet und deren Auswirkungen auf die Spielstärke und Performance untereinander verglichen werden.

Vorläufige Gliederung

1. Einleitung
    1.1 Motivation
    1.2 Ziele und Forschungsfrage
    1.3 Aufbau der Arbeit
2. Grundlagen
    2.1 Reinforcement Learning
        2.1.1 Begriffserklärung und Verortung
        2.1.2 Markov Decision Process
        2.1.3 Dynamic Programming
        2.1.4 Temporal Difference Learning
        2.1.5 Exploitation vs Exploration
    2.2 Q-Learning
    2.3 SARSA
    2.4 MiniMax-Algorithmus
    2.5 Tic-Tac-Toe
        2.5.1 Spielerklärung
        2.5.2 Anwendbare Reinforcement Learning Algorithmen
3. Methodik und Funktionsweise der Algorithmen
    3.1 Spielfeld
    3.2 Agent Q-Learning
    3.3 Agent SARSA
    3.4 Trainingsaufbau und Evaluationsmetriken
4. Implementierung
5. Diskussion und Auswertung der Ergebnisse
    5.1 Auswertung des Q-Learning Agenten
    5.2 Auswertung des SARSA Agenten
6. Konklusion
    6.1 Beantwortung der Forschungsfragen
    6.2 Kritische Betrachtung der Inhalte
    6.3 Anmerkungen für künftige Arbeiten

Literaturverzeichnis

[1] L. V. Allis, „Searching for solutions in games and artificial intelligence,“ Ponsen & Looijen, Wageningen, 1994, ISBN: 9789090074887.

[2] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Pr, 1957, 339 S., ISBN: 978-0-691-07951-6.

[3] R. Bellman, „The theory of dynamic programming,“ *Bulletin of the American Mathematical Society*, Jg. 60, Nr. 6, S. 503–515, 1954. DOI: 10.1090/S0002-9904-1954-09848-8.

[4] M. Bowling, „Convergence and No-Regret in Multiagent Learning,“ University of Alberta Libraries, 2004. DOI: 10.7939/R3ZS2KF41.

[5] J. A. Boyan, „Modular Neural Networks for Learning Context-Dependent Game Strategies,“ Master's thesis, Computer Speech and Language Processing, 1992.

[6] W. E. Deming, J. von Neumann und O. Morgenstern, „Theory of Games and Economic Behavior.,“ *Journal of the American Statistical Association*, Jg. 40, Nr. 230, S. 263, Juni 1945, ISSN: 01621459. DOI: 10.2307/2280142. JSTOR: 2280142.

[7] S. L. Epstein, „Toward an ideal trainer,“ *Machine Learning*, Jg. 15, Nr. 3, S. 251–277, Juni 1994, ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00993346.

[8] W. Ertel, *Introduction to Artificial Intelligence* (Undergraduate Topics in Computer Science). Cham: Springer International Publishing, 2017. DOI: 10.1007/978-3-319-58487-4.

[9] R. A. Howard, *Dynamic Programming and Markov Processes*, 6. print. Cambridge, Mass: M.I.T. Pr, 1960, 136 S., ISBN: 978-0-262-08009-5.

[10] W. Konen und T. Bartz–Beielstein, „Reinforcement Learning: Insights from Interesting Failures in Parameter Selection,“ in *Parallel Problem Solving from Nature – PPSN x*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, S. 478–487, ISBN: 978-3-540-87700-4.

[11] M. L. Littman, „Markov Games as a Framework for Multi-Agent Reinforcement Learning,“ in *In Proceedings of the Eleventh International Conference on Machine Learning*, Morgan Kaufmann, 1994, S. 157–163. DOI: 10.1.1.48.8623.

[12] G. Neto, „From Single-Agent to Multi-Agent Reinforcement Learning: Foundational Concepts and Methods,“

[13] G. Rummery und M. Niranjan, „On-Line Q-Learning Using Connectionist Systems,“ *Technical Report CUED/F-INFENG/TR 166*, 4. Nov. 1994.

[14] S. J. Russell und P. Norvig, *Artificial Intelligence: A Modern Approach* (Pearson Series in Artificial Intelligence), Fourth edition. Hoboken: Pearson, 2021, ISBN: 978-0-13-461099-3.

[15] M. A. Samsuden, N. M. Diah und N. A. Rahman, „A Review Paper on Implementing Reinforcement Learning Technique in Optimising Games Performance," in *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*, Shah Alam, Malaysia: IEEE, Okt. 2019, S. 258–263, ISBN: 978-1-72810-758-5. DOI: 10.1109/ICSEngT.2019.8906400.

[16] R. S. Sutton, „Learning to Predict by the Methods of Temporal Differences," *Machine Learning*, Jg. 3, Nr. 1, S. 9–44, 1988, ISSN: 08856125. DOI: 10.1023/A:1022633531479.

[17] R. Sutton und A. Barto, „Toward a Modern Theory of Adaptive Networks: Expectation and Prediction.," *Psychological review*, Jg. 88, Nr. 2, S. 135–170, 1981. DOI: 10.1037/0033-295X.88.2.135.

[18] R. S. Sutton und A. G. Barto, *Reinforcement Learning: An Introduction* (Adaptive Computation and Machine Learning Series), Second edition. Cambridge, Massachusetts: The MIT Press, 2018, 526 S., ISBN: 978-0-262-03924-6.

[19] C. Szepesvari, *Algorithms for Reinforcement Learning* (Synthesis Lectures on Artificial Intelligence and Machine Learning 9). San Rafael, Calif.: Morgan & Claypool, 2010, 89 S., ISBN: 978-1-60845-492-1.

[20] I. Szita, „Reinforcement Learning in Games," in *Reinforcement Learning*, Ser. Adaptation, Learning, and Optimization, Bd. 12, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, S. 539–577, ISBN: 978-3-642-27645-3. DOI: 10.1007/978-3-642-27645-3_17.

[21] G. Tesauro, „Temporal Difference Learning of Backgammon Strategy," in *Machine Learning Proceedings 1992*, Elsevier, 1992, S. 451–457, ISBN: 978-1-55860-247-2. DOI: 10.1016/B978-1-55860-247-2.50063-2.

[22] C. J. C. H. Watkins, „Learning From Delayed Rewards," 1. Jan. 1989.

[23] C. J. C. H. Watkins und P. Dayan, „Q-learning," *Machine Learning*, Jg. 8, Nr. 3, S. 279–292, 1. Mai 1992, ISSN: 1573-0565. DOI: 10.1007/BF00992698.