

**<NAZWA WYDZIAŁU>**

Praca dyplomowa

*Zastosowanie uczenia ze wzmocnieniem w celu zastąpienia regulatora instalacji grzewczych*

*Reinforcement Learning Applications to Control of Heating Systems*

Autor: *Adrian Gacek*

Kierunek studiów: Automatyka i Robotyka

Opiekun pracy: *dr inż. Jakub Grela*

Kraków, 2024

# Wstęp

W ciągu ostatnich lat, można zaobserwować znaczące przyśpieszenie rozwoju sztucznej inteligencji, a jej potencjał znajduje zastosowanie w coraz szerszym zakresie problemów. Od generowania obrazów po zarządzanie ruchem drogowym i identyfikację celów militarnych, algorytmy uczenia maszynowego odgrywają kluczową rolę w rozwiązywaniu zarówno prostych, jak i niezwykle złożonych wyzwań. Jednym z obszarów, w którym sztuczna inteligencja staje się coraz bardziej znacząca, jest skuteczne i zoptymalizowane sterowanie systemami grzewczymi.

Współczesne systemy grzewcze odgrywają istotną rolę w zapewnieniu komfortu i bezpieczeństwa w budynkach, a ich odpowiednia regulacja jest kluczowa dla efektywnego wykorzystania energii. Ze względu na powszechność tego problemu oraz duże znaczenie ogrzewania w kontekście kosztów eksploatacji budynku, nawet niewielkie poprawy w sterowaniu mogą mieć istotny wpływ zarówno na budżet, jak i na środowisko.

W świetle powyższego, niniejsza praca ma na celu przeprowadzenie analizy możliwości wykorzystania algorytmów uczenia ze wzmocnieniem w problemach sterowania systemami grzewczymi. Użycie tych zaawansowanych technik może prowadzić do bardziej efektywnego wykorzystania energii, poprawy komfortu użytkowników oraz redukcji negatywnego wpływu na środowisko naturalne.

Kolejne rozdziały będą miały na celu kolejno przedstawienie teoretycznych podstaw zagadnienia uczenia maszynowego, w szczególności przegląd metod uczenia ze wzmocnieniem i istotnych pojęć z nimi związanych, oraz zaprezentowane zostaną przykłady zastosowania podobnych rozwiązań w literaturze. Następnie w części praktycznej przedstawiony zostanie proces przygotowania środowiska opartego o symulację TCLab i uczenia wskazanych w dalszych rozdziałach algorytmów. Na koniec nakreślono wyniki testów symulacyjnych oraz sprzętowych, wraz z wnioskami.

# Podstawy teoretyczne

Niniejszy rozdział ma za zadanie przedstawić podstawowe idee stojące za uczeniem maszynowym i metodami uczenia ze wzmocnieniem.

### Uczenie maszynowe

Uczenie maszynowe to obszerna dziedzina sztucznej inteligencji, zajmująca się rozwijaniem algorytmów i technik umożliwiających komputerom uczenie się z danych i doświadczenia, bez konieczności jawnego programowania. Polega ono na tworzeniu modeli, które analizując dane, wyszukują wzorców i podejmują decyzje lub prognozy na ich podstawie.

Uczenie maszynowe można podzielić na następujące obszary:

* Uczenie nadzorowane
* Uczenie nienadzorowane
* Uczenie ze wzmocnieniem

Algorytmy uczenia nadzorowanego otrzymują zestaw danych wejściowych, w którym każdy przykład posiada etykietę określającą pożądane wyjście. Celem jest znalezienie funkcji mapującej dane wejściowe na dane wyjściowe. Są one stosowane na przykład w problemach klasyfikacji i regresji.

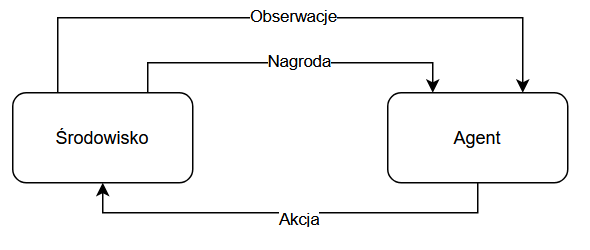
Uczenie nienadzorowane zakłada, że algorytm otrzymuje dane wejściowe bez etykiet i ma za zadanie znajdować wzorce występujące w danych. Taki rodzaj algorytmów może być zastosowany do klasteryzacji, czy wykrywania anomalii.

### Uczenie ze wzmocnieniem

### Uczenie ze wzmocnieniem (w skrócie RL, z angielskiego Reinforcement Learning) to proces uczenia się, w którym celem jest maksymalizacja nagrody poprzez przyporządkowanie sytuacji do akcji. Uczący się algorytm nie otrzymuje wskazówek, które akcje wykonać, lecz nabiera doświadczenia poprzez przeprowadzenie interakcji z systemem. Tworzy się w ten sposób model wskazujący, które akcje przynoszą największą nagrodę, często biorąc pod uwagę nie tylko stan aktualny, ale również następujące po nim stany przyszłe. Poszukiwanie metod prób i błędów oraz opóźniona nagroda to dwie główne cechy odróżniające uczenie ze wzmocnieniem od innych form uczenia. Problematyka uczenia ze wzmocnieniem jest formalizowana przy użyciu teorii systemów dynamicznych. [1]

### Agent i środowisko

Proces uczenia ze wzmocnieniem opiera się o interakcję między agentem a środowiskiem: agent wykonuje akcję, na którą reaguje środowisko, zmieniając swój stan i przyznając za to nagrodę. Następnie agent otrzymuje nagrodę oraz obserwacje zawierające informacje o stanie środowiska, które wykorzystuje do dalszej nauki i optymalizacji swoich przyszłych decyzji. Celem jest opracowanie strategii, która maksymalizuje sumę przyszłych nagród, które agent może otrzymać.



Rys. 1. Interakcja między środowiskiem i agentem. Opracowanie własne przy pomocy narzędzi draw.io

### Podstawowe pojęcia w RL

Dla pełnego zrozumienia procesu uczenia ze wzmocnieniem należy zdefiniować jego podstawowe elementy [1]:

**Obserwacje** – są to wszystkie informacje o stanie środowiska, które umożliwiają określenie następnej akcji podjętej przez agenta. Warto zauważyć, że nie w każdym systemie agent ma dostęp do wszystkich zmiennych stanu, co może utrudnić proces uczenia.

**Akcja** – może być zdefiniowana w sposób zarówno ciągły, jak i dyskretny. Umożliwia wpływanie agenta na zmianę stanu środowiska.

**Nagroda** - zmienna zwracana przez środowisko po wykonaniu kroku przez agenta, obrazuje to, jak korzystne było wykonanie akcji. Celem agenta jest maksymalizowanie sumarycznej nagrody gromadzonej podczas całego cyklu życia procesu.

**Polityka –** funkcja mapująca obserwacje ze środowiska na akcje. Ma za zadanie zdefiniować zachowanie agenta w sytuacjach, w jakich może znaleźć się środowisko. Znalezienie odpowiedniej polityki jest celem uczenia ze wzmocnieniem. Może być definiowana zarówno deterministycznie, jak i stochastycznie.

**Funkcja wartości** - określa, co jest korzystne w dłuższej perspektywie czasowej, w przeciwieństwie do sygnału nagrody, który ocenia tylko natychmiastowe efekty akcji. W przybliżeniu wartość stanu to całkowita suma nagród, jakiej agent może oczekiwać, zaczynając od danego stanu. Ma na celu wskazać na długoterminową atrakcyjność stanów, biorąc pod uwagę stany, które prawdopodobnie nastąpią oraz nagrody dostępne w tych stanach.

**Stopę dyskontową** – współczynnik określający relację między natychmiastową nagrodą, a przewidywaną sumą przyszłych nagród. Wzór (1.1) wskazuje typowe zastosowanie:

|  |  |  |
| --- | --- | --- |
|  |  | (1.1) |

Gdzie:

* – maksymalizowana w procesie uczenia wartość sumy nagród oceniająca korzyść z wykonania danej akcji
* numer stanu, dla którego wyznaczamy wartość sumy nagród
* – stopa dyskontowa, wartość pomiędzy zapewnia, że każdy następny przewidywany stan, ma mniejsze znaczenie względem natychmiastowej nagrody
* – numer przyszłego przewidywanego stanu
* – Przyszła przewidywana nagroda

### Eksploracja, a eksploatacja

Istotnym problemem w dziedzinie RL jest problem balansu między eksploracją i eksploatacją. Wynika on z faktu, że zbieranie informacji i ich wykorzystywanie to często dwa wzajemnie wykluczające się działania. Z jednej strony, zależy nam, aby agent eksplorował, czyli zdobywał informacje o swoim otoczeniu w celu poprawy przyszłych wyników. Ponosi jednak w ten sposób koszty zdobywania nowych informacji nie korzystając z posiadanej wiedzy. Z drugiej strony, eksploatując czyli używając wyłącznie istniejącej wiedzy, ograniczamy zdolności adaptacji do istotnych zmian w otoczeniu i system może utknąć w suboptymalnej stabilnej równowadze. Dlatego optymalne zachowanie zwykle wymaga pewnej równowagi między działaniami eksploracyjnymi a eksploatacyjnymi. [2]

Przykładem rozwiązania tego problemu może być jedna z prostych metod balansowania między eksploracją a eksploatacją ε-greedy. Działa ona w następujący sposób [3]:

* **Eksploatacja**: Większość czasu metoda ε-greedy działa zachłannie (greedy), czyli wybiera działanie, które w danej chwili wydaje się najlepsze na podstawie dostępnej wiedzy.
* **Eksploracja**: Co pewien czas, z małym prawdopodobieństwem ε, metoda wybiera losowo jedno z dostępnych działań, zamiast wybierać działanie uważane za najlepsze.

# Bibliografia

1. Sutton, R. S., & Barto, A. G. (2018). "Reinforcement Learning: An Introduction." <http://incompleteideas.net/book/RLbook2020.pdf>
2. Berger-Tal O, Nathan J, Meron E, Saltz D (2014) The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. PLoS ONE 9(4): e95693. <https://doi.org/10.1371/journal.pone.0095693>
3. Alexandre dos Santos Mignon, Ricardo Luis de Azevedo da Rocha, An Adaptive Implementation of ε-Greedy in Reinforcement Learning, Procedia Computer Science, Volume 109, 2017, Pages 1146-1151, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.05.431>. (https://www.sciencedirect.com/science/article/pii/S1877050917311134)