Kursinis darbas

Žaidimų žaidimas naudojantis skainamuoju mokymu

Game playing using reinforcement learning

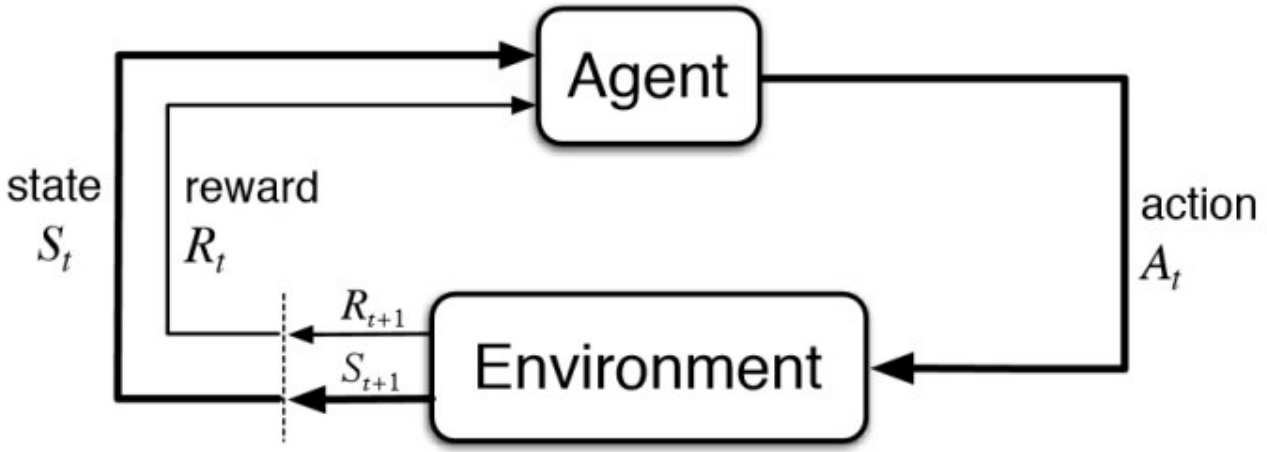
Arnoldas Čiplys

Darbo tikslas

- Išanalizuoti, aptarti ir palyginti skatinamojo mokymo algoritmus
- Išanalizuoti kaip skatinamasis mokymas naudojamas modernių žaidimų agentams mokyti

Skatinamasis mokymas

Mašininio mokymosi sritis, kurioje sprendžiama kaip agentas turi elgtis duotoje aplinkoje, kad gautų didžiausią atlygį.



Klasikiniai algoritmai

- Monte Karlo metodai
- Q-mokymasis (angl. Q-learning)
- Temporal-Difference mokymas

Q-mokymasis (angl. Q-learning)

1. Sukuriame matricą (q-lentelę), kurios dydis yra būsenų kiekis x veiksmai ir visus langelius priskiriame 0 (angl. q-table)
2. Agentas pasirenka veiksmą:
   1. Pasirenkame geriausią veiksmą iš q-lentelės pagal dabartinę būseną
   2. Judame betkuria kryptimi, kad atrastume naujų būsenų, kurių kitu atvejų galbūt nepasiektume
3. Atnaujiname q-lentelę pagal formulę:

$$Q^{new}(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\Big(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \Big)}^{\text{temporal difference}}$$

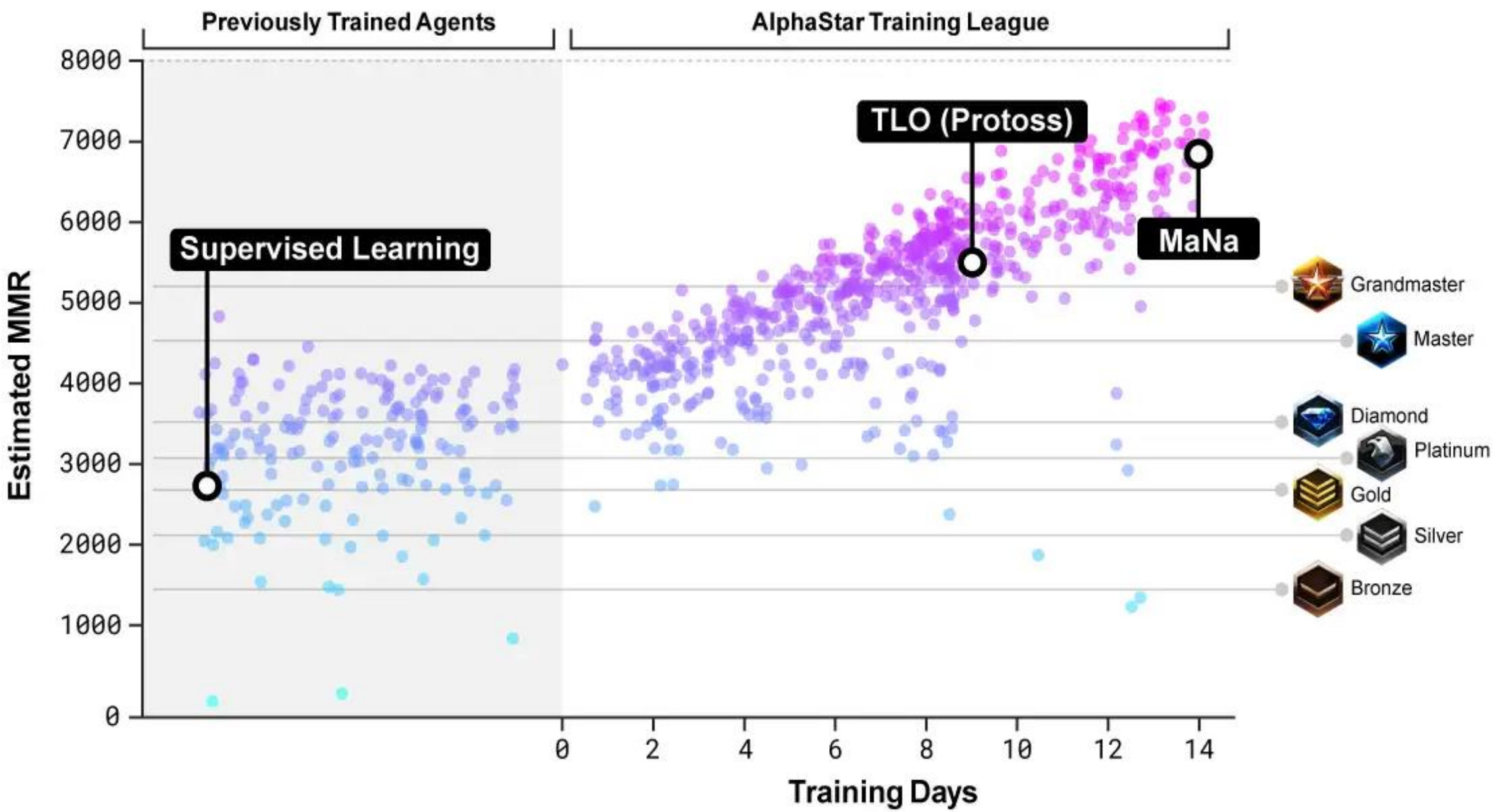
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{new value (temporal difference target)}}$$

Modernūs algoritmai

- AlphaStar (Starcraft 2)
- OpenAI (Dota 2)
- AlphaZero (šachmatai)

AlphaStar

- Pradiniam mokymui naudotas prižiūrimas mokymas (angl. supervised learning)
- Agento elgesys nusprendžiamas naudojant gilųjį neuroninį tinklą (angl. deep neural network)
- Agento neuroninis tinklas treniruojamas naudojant multi-agentų (angl. multi-agent) skatinamojo mokymo algoritmą, kuriame agentas žaidžia prieš savo kopijas


Previously Trained Agents
AlphaStar Training League
Estimated MMR
8000
7000
6000
5000
4000
3000
2000
1000
0
Supervised Learning
TLO (Protoss)
MaNa
Grandmaster
Master
Diamond
Platinum
Gold
Silver
Bronze
0
2
4
6
8
10
12
14
Training Days

Išvados

- Skatinamasis mokymas nėra labai efektyvus, jei aplinkoje yra labai daug skirtingų būsenų ir galimų veiksmų
- Skatinamojo mokymo algoritmai turi skirtingus bruožus, pagal kuriuos galime nuspresti ar jis tinkamas atliekamai užduočiai (pvz. modelis, strategija, veiksmo erdvė)
- Modernūs skatinamojo mokymo taikymai kombinuoja jį kartu su skirtingais mašininio mokymo algoritmais, kad minimizuotų skatinamojo mokymo trūkumus