RL в покере и играх с неполной информацией

25 мая

Шошин Борис

В чем отличие?

В играх с неполной информацией не существует однозначно определенного правильного решения. Оптимальным будет определенное вероятностное распределение между действиями, которое нам необходимо определить.

Пример: камень-ножницы-бумага

В данной игре оптимально выбирать каждый из вариантов с вероятностью ⅓, при этом любая другая стратегия будет легко эксплуатируемой.

При этом в данной ситуации состояние при котором оба игрока выбирают все действия с равной вероятностью называется равновесием Нэша

Равновесие Нэша

Равновесие Нэша - набор стратегий, при котором ни один участник не может увеличить выигрыш, изменив свою стратегию, если другие участники своих стратегий не меняют.

Существует для любых конечных игр с любым количеством игроков.

Наш идеал - найти данную стратегию.

Regrets

Regret - мера того, как сильно мы сожалеем, что выбрали данное действие относительно какого-то другого

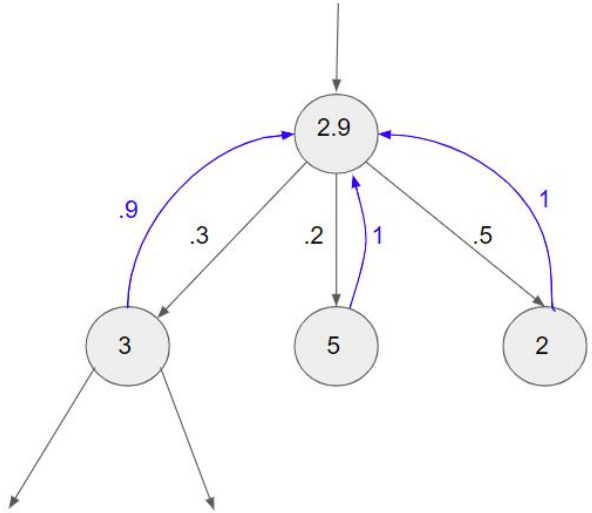
regret = u(possible action) - u(action taken)

	R	P	S
R	$0, 0$	$-1, 1$	$1, -1$
P	$1, -1$	$0, 0$	$-1, 1$
S	$-1, 1$	$1, -1$	$0, 0$

Payoff grid of Rock Paper Scissors game

Counterfactual Regret Minimization

Будем рассматривать игры с несколькими шагами и считать значения вершин (counterfactual value) рекурсивно.

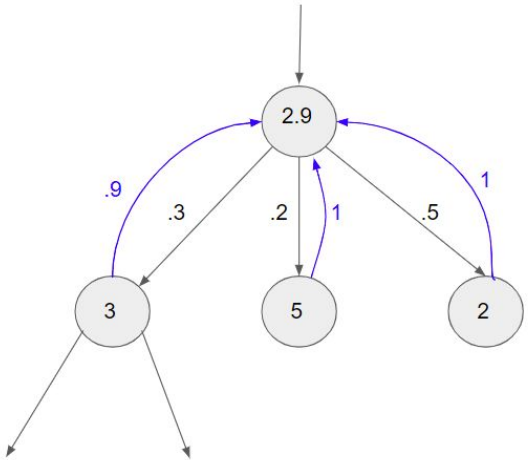


После этого мы можем посчитать все regrets и на основании их обновить свою стратегию. Для этого для каждой вершины помним их суммарные сожаления для каждого действия и прибавляем текущее сожаления.

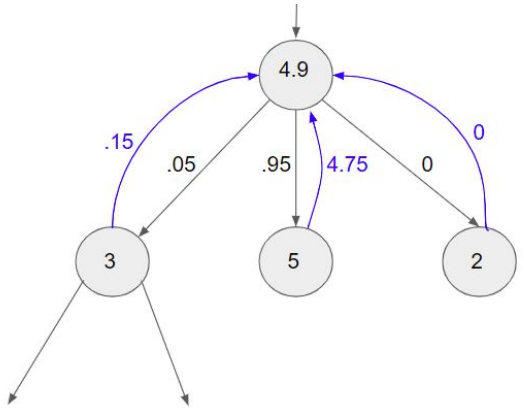
Обновление стратегии

В качестве стратегии будем просто брать вероятности, пропорциональные величинам накопленных сожалений. Будем считать все сожаления неотрицательными.

$$p_i = \frac{regret_i}{\sum regret_j}$$

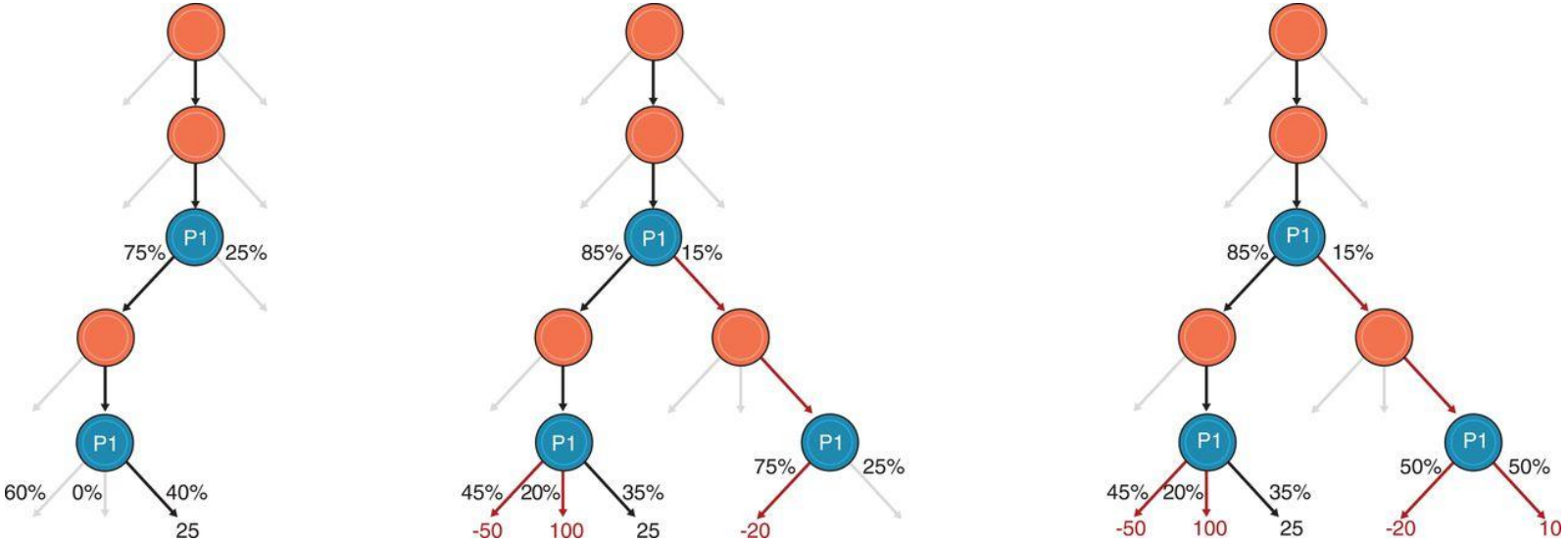


$r_1 = 0.1$
$r_2 = 2.1$
$p_1 = 0.1 / 2.2$
$p_2 = 2.1 / 2.2$



MCCRF

Делаем случайную симуляцию, после чего из всех точек принятия решения также делаем случайную симуляцию и считаем сожаления. Затем повторяем то же самое для точек принятия решения из предыдущих симуляций.



Self-play

Для обучения этого используется self-play. Инициализируем веса случайно и дальше обновляем случайно, после чего начинаем играть сами с собой и выбирать случайно одну из своих предыдущих версий в качестве соперника.

Pluribus

Для достижения хороших результатов необходимо много абстракций, которые нужно задавать руками для каждой игры.

Использует два типа абстракций:

Абстракции действия: используются для уменьшения количества возможных действий. Так, ставка в 200$ и в 201$ будут восприниматься как одно и то же действие

Информационная абстракция: используется, для лучшего понимания игры, так стрит начинающийся с 9 и начинающийся с 10 будет рассматриваться как одна и та же комбинация

Pluribus

Состоит из 2 частей:

1. Offline обученная стратегия.
2. Online depth-limited search

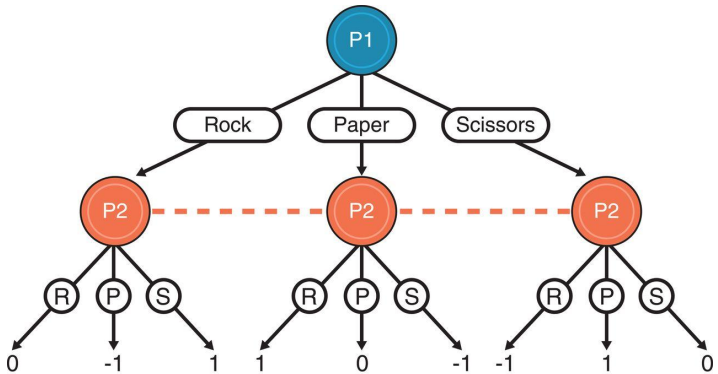
Offline стратегия

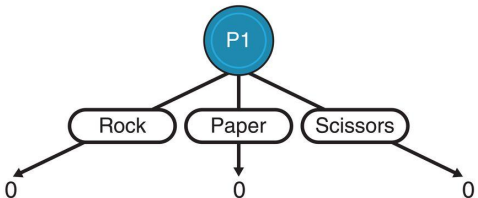
Предобученное на self-playing MCCFR с небольшими изменениями. Regrets будет добавляться линейно растущий коэффициент. Таким образом, мы будем обращать внимание на опыт полученный в начале при очень случайных стратегиях игры.

При игре в покер используется при выборе первых действий(на префлопе).

Online поиск

В играх с неполной информацией нельзя использовать обычный поиск в глубину из-за того, что он будет давать не распределение вероятностей, а всегда одно и то же действие.





Online поиск

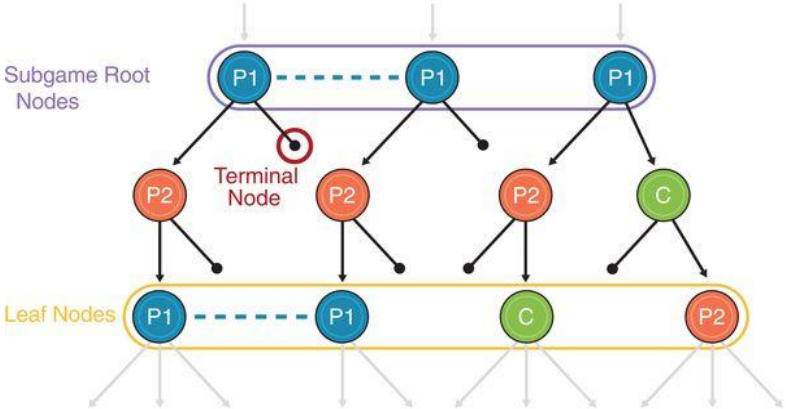
Для решения этой проблемы берется определенное k стратегией и мы считаем, что соперник может выбрать одну из них. Pluribus использует 4 стратегии. Одна из них - стратегия обученная MCCFR, а три другие - её модифицированные варианты в сторону большего количества пассов / коллов / рейзов.

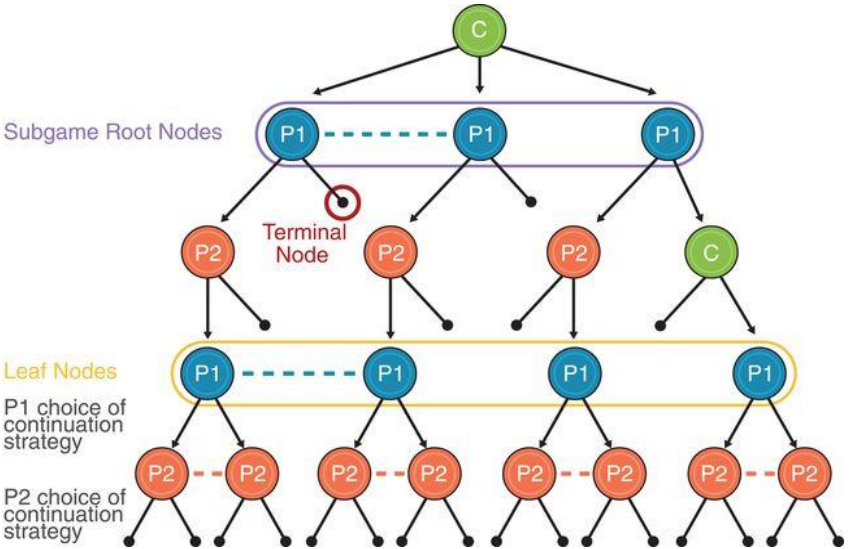
Online поиск

Вторая большая проблема - оптимальная стратегия зависит также от того как мы будем играть в данной ситуации с другими картами. Для решения. Поэтому pluribus считает вероятность того, что он окажется в данной ситуации со всеми другими руками.

Online поиск

Таким образом при выборе действия мы достоверно не знаем в какой именно ситуации находимся. При выборе действия мы делаем поиск в глубину, и после того как выбрали стратегию для каждого игрока осуществляем выбор оптимального действия с помощью либо MCCRF, либо CRF в зависимости от стадии игры.





Источники

- https://www.cs.cmu.edu/~noamb/papers/17-IJCAI-Libratus.pdf
- https://papers.nips.cc/paper/2012/file/3df1d4b96d8976ff5986393e8767f5b2-Paper.pdf
- http://poker.cs.ualberta.ca/publications/NIPS07-cfr.pdf
- https://science.sciencemag.org/content/365/6456/885
- https://towardsdatascience.com/counterfactual-regret-minimization-ff4204bf4205