Event-Driven Models
Събитийни модели

Dimiter Dobrev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
d@dobrev.com

При Reinforcement Learning търсим смисъл в потока входно-изходна информация. Ако не намерим смисъл, този поток за нас ще бъде просто шум. За да намерим смисъл трябва да се научим да откриваме и разпознаваме обекти. Какво е обект? В тази статия ще покажем, че обекта е event-driven модел. Тези модели са обобщение на action-driven моделите. При Markov decision process имаме action-driven модел и там състоянието на модела се променя на всяка стъпка. Предимството на event-driven моделите е, че те са по-устойчиви и променят състоянието си само при настъпването на някакви определени събития. Тези събития може да се случват много рядко и затова състоянието на event-driven моделите е много по-предвидимо.

Keywords: Artificial Intelligence, Reinforcement Learning, Partial Observability, Event-Driven Model, Action-Driven Model, Object.

Въведение

Какво е обект? Пример за обект може да бъде вашата любима песен, определена дума или синтактична категория (като глагол), това може да е човек, предмет или животно, това може да е вашата къща или квартала, в който живеете.

Вие можете да разпознавате обекти. Вашата любима песен ще я познаете дори и да чуете само част от нея. Дори и певеца да е друг, дори и качеството да е лошо, вие пак ще разпознаете песента.

Обектите в главата ви са подредени йерархично. Пример за йерархия: животно, куче, пудел, вашето куче. Някой обекти притежават свойството единственост. Например вашето куче е единствено и ако го подстрижете, то то ще е подстригано. Ако подстрижете произволно куче, от там няма да следва, че всички кучета ще бъдат подстригани.

Свойството единственост няма да е характеристика, която да не може да се промени. Например смятате един човек за единствен, но установявате, че той има брат близнак. Мислите, че сте говорили с един и същи човек, но се оказва че сте говорили с двама различни. Друг пример. Във вашата къща имате стол. Смятате, че този стол е единствен. Знаете, че са произведени хиляди такива столове, но този е единствения такъв стол, който имате вкъщи. Боядисвате стола и очаквате следващия път като го видите той да е боядисан. Може да се окаже, че във вашата къща има два такива стола, а вие сте боядисали само единия.

Как разпознаваме обектите? Вашата любима песен може да я чуете от начало, от средата или от края. Човека можете да го видите в профил или в анфас. Квартала може да го видите тръгвайки от много различни кръстовища. Тоест, за да разпознаем един обект, на нас не ние е нужно да го видим по съвсем същия начин, по който вече сме го виждали.

Ориентираният граф също е обект и ние го разпознаваме когато минем определен път в този граф. Разбира се, пътя трябва да е достатъчно характерен, за да сме сигурни, че това е точно този ориентиран граф. Ако не сме сигурни, може да направим експеримент и да завием в тази посока, която ще ни даде повече информация. Например, вие сте на едно кръстовище, но не знаете дали сте във вашия квартал. Тръгвате към главната улица, за да се ориентирате. Друг пример. Виждате предмет, който не можете да разпознаете. Започвате да го въртите в ръцете си и да го оглеждате. Виждате познат човек, но не сигурни, че е той. Започвате да го оглеждате, викате го по име, за да видите дали ще се обърне.

Не винаги имаме контрола и не винаги можем да направим експеримент. Когато чуем една дума понякога не можем да разберем какво сме чули. Може да имаме възможност да помолим да ни повторят думата или ако е на запис може да върнем записа и да я чуем отново, но в повечето случай се получава: който разбрал – разбрал.

В тази статия обектите ще ги представяме като ориентирани графи (event-driven модели). Повечето обекти не са ни постоянно пред очите. Тоест, обектите се появяват и изчезват от нашето полезрение. Има ли обекти, които наблюдаваме постоянно. Има ли ориентирани графи, в които ние се намираме постоянно и никога не ги напускаме. Има такива обекти (модели). Например дните от седмицата. Имаме един модел със седем състояния (дните от седмицата) и ние постоянно сме в едно от тези седем състояния. Друг пример е нашия квартал. Ако цял живот никога не излезем от нашия квартал, то това ще е обект, който винаги е в полезрението ни.

Повечето event-driven модели ще имат едно специално състояние, което ще наречем *outside*. Ще считаме, че обекта е извън нашето полезрение, когато сме в състоянието *outside*.

При event-driven моделите основният въпрос ще бъде „Къде съм?“ Тоест, в кое състояние съм? Ако event-driven модела е обект, тогава основният въпрос е „Виждам ли обекта?“ Тоест, дали съм в състоянието *outside* или в някое от другите състояния.

Тази статия ще я започнем представяйки интуитивна идея за event-driven моделите. После ще направим сравнение между тези и action-driven моделите. Ще опишем задачата неформално и ще кажем защо няма да търсим началното състояние. Ще дадем дефиниция на история и ще кажем каква е целта на агента. Ще покажем, че можем да имаме много различни критерии и че Sutton в [2] не е прав като изказва, че има само един критерии. Ще покажем, че понятието discount factor не трябва да участва в дефиницията на Reinforcement Learning (RL), защото той е свързан със стратегията, а не със смисъла на живота. Ще обсъдим дали модела на света трябва да е детерминиран или да не е. За некоректните ходове ще покажем, че можем да минем и без тях, но е по-добре да предполагаме, че има такива. Ще дадем дефиниция на RL. После ще дефинираме съвършения модел, който е action-driven. Ще усложним този модел като добавим случайност. Ще покажем, че трябва да има отпечатък, тоест трябва да се случва нещо специално, което да отличава различните

състояния. Ще въведем пълните модели и ще покажем, че те са също така непостижими като съвършените. Ще кажем какво е събитие и ще направим следващото усложняване на модела, което ще го превърне от action-driven в event-driven. Тоест, ще заменим действията с произволни събития. Ще добавим към модела и променливи и ще направим декартовото произведение от всички адекватни модели, които сме открили. Това декартово произведение ще бъде модела, който търсим и който възможно най-добре описва света.

Интуитивна идея

Преди да сме объркали читателя с много приказки ще се опитаме да дадем интуитивна идея за това какво представлява event-driven модела. Това е ориентиран граф подобен на фигури 1 и 2.

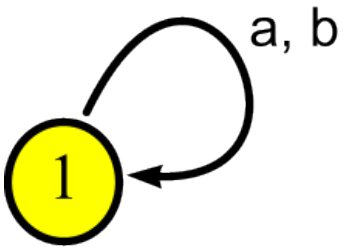

Figure 1

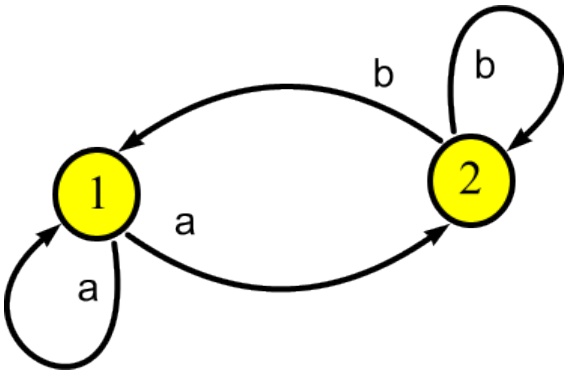

Figure 2

Тук *a* и *b* са събития. Ако *a* и *b* бяха действия, то тогава фигури 1 и 2 щяха да изобразяват action-driven модели. Разбира се, действието е събитие и затова action-driven моделите са частен случай на event-driven моделите.

Модела на фигура 1 е много елементарен и ако направим статистика по него, единственото което ще отчетем е колко пъти се е случило събитието *a* и колко пъти събитието *b*. Тоест, ще имаме идея кое събитие е по-вероятно да се случи.

Модела на фигура 2 е по-интересен. Тук, ако сме в състоянието 1 събитието *b* не може да се случи. Същото можем да кажем за състоянието 2 и събитието *a*. Тоест, този модел ни предсказва кое ще е следващото събитие (*a* или *b*). Ако знаем в кое състояние сме, ще знаем кое ще е следващото събитие.

Ако обърнем стрелките (фигура 3), ще получим модел, който помни кое е последното събитие (*a* или *b*). Ако знаем в кое състояние сме, ще знаем кое събитие последно се е случило.

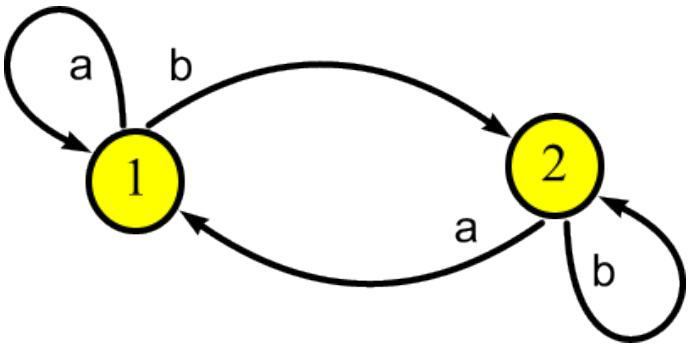

Figure 3

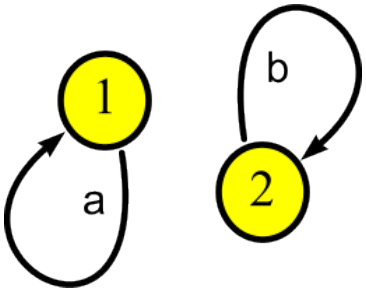

Figure 4

Без значение дали разглеждаме фигура 2 или 3, и в двата случая, трябва да има разлика между състоянията 1 и 2. Трябва нещо специално да се случва в някое от тези състояния. Това нещо ще го наречем отпечатък. Ако няма отпечатък, т.е. ако няма нищо специално, то тогава двата модела са безсмислени.

При фигура 2 има разлика между състояния 1 и 2 и тя е, че в 1 излизат стрелки само по a (и в 2 само по b). Ако няма никаква друга разлика между 1 и 2, тогава няма да можем да кажем в кое състояние сме, защото тези състояния ще са напълно неотличими спрямо миналото.

При фигура 3 знаем точно в кое състояние сме, но това няма никакъв смисъл, ако тези състояния са напълно неотличими спрямо бъдещето. Тоест, няма да има никакъв смисъл да помним дали последно се е случило a или b, ако от това не следва нищо за бъдещето.

Като пример нека да разгледаме следната криминална история (фигура 4). Търсим кой е убиеца. Дали е пощальона или скитника? Ако убиеца е пощальона, тогава сме в състояние 1. Ако убиеца е скитника, тогава сме в състояние 2. Събитието a е „доказахме, че убиеца е пощальона“. Съответно събитието b е „доказахме, че убиеца е скитника“. Въпросът е, в кое състояние сме? Тоест, кой е убиеца? Когато се появи доказателството (събитието a или b) ще разберем. За съжаление, доказателството може и никога да не се появи. Въпросът е, как да разберем кой е убиеца преди да се е появило доказателството. Може да предположим, че когато убиецът е пощальона е по-вероятно да се появи улика потвърждаваща това, отколкото улика отричаща това. Тоест, уликите не доказват кой е убиеца, но ни дават кое от двете е с по-голяма вероятност.

Модела на фигура 4 на нас не ни харесва защото ние искаме да имаме повторяемост и по всяка стрелка да сме минали многократно. В този модел, ако се е случило a, то тогава b няма как да се случи.

Повторяемостта е много важна. Хераклит е казал, че човек не може да влезе два пъти в една и съща река. Реката всеки път е различна. По тази логика не можете да влезете два пъти в една и съща стая и не можете да срещнете два пъти един и същи човек. Стаята всеки път е различна. Някой е боядисал стените, друг е разместил мебелите. Един е влязъл, друг е излязъл. Ако всичко е по старому, ще се появи някоя муха, която ще попречи на повторяемостта.

Въпреки всичко, ние искаме да влизаме много пъти в една и съща река, и в една и съща стая и да срещаме един и същи човек. Те може да не са съвсем едни и същи, но ние сме готови да пренебрегнем малките разлики и дори и големите разлики, но да направим света по-прост и по-разбираем.

Това е идеята която стои зад event-driven моделите. Това са модели, при които имаме малко състояния, които посещаваме многократно и всяко наше посещение е дълго (състои се от много стъпки). Обратното е при action-driven моделите, особено ако модела е съвършен (виж по-долу). Съвършен модел винаги съществува и най-лесния начин да направим такъв модел е да наредим състоянията в редица и да минем през тях еднократно. В съвършеният модел състоянието описва всичко и затова е почти невъзможно едно и също състояние да се повтори два пъти (както не можем да влезем два пъти в една и съща река).

Event-Driven vs Action-Driven

Каква е разликата между event-driven и action-driven моделите. При Markov decision process (MDP) модела е един ориентиран граф, който променя състоянието си след всяко действие (т.е. на всяка стъпка). Такъв модел е като машина, която щрака твърде бързо. Много трудно е при такъв модел да си отговорим на въпроса „Къде съм?“. Тоест, много трудно е да кажем кое е текущото състояние. Да си представим, че човекът е стъпково устройство, което прави 24 стъпки в секунда. (В киното кадъра се сменя 24 пъти в секунда и ние приемаме действието като непрекъснато. Тоест, 24 стъпки в секунда е едно добро предположение.) Ако текущото ви състояние се променя 24 пъти в секунда, едва ли бихте ли могли да кажете кое е текущото ви състояние. Нека вземем модел, който не се променя при всяко действие, а при появата на някакви по-интересни събития. Пример за такива събития са изгрева и залеза. Те се случват по веднъж в денонощието. След изгрева от нощ преминаваме към ден, а след залеза обратното. Ако ви попитам дали в момента е ден или нощ вие вероятно ще успеете да ми отговорите. Тоест, текущото състояние на event-driven модела е много по-предвидимо (защото е много по-стабилно и по-рядко се променя).

Друга разлика е, че при action-driven моделите може да се опитаме да опишем поведението на света без това да включва поведението на агента докато при event-driven моделите се налага да описваме света и агента живеещ в него като една единна система.

Например при MDP изчисляваме каква е вероятността за определен преход при определено състояние и определено действие. Ние обаче не изчисляваме каква е вероятността агента да извърши определеното действие, защото се опитваме да опишем само света без да описваме агента. При event-driven моделите се налага да опишем света и агента като единна система, защото когато се е случило едно събитие ние не можем да кажем, дали това е само по волята на агента или това се е случило, защото светът е такъв какъвто е. При MDP ние не предсказваме миналото, защото по коя стрелка сме влезли в едно състояние зависи не само от света, но и от агента. Например (фигура 5), ако сме в състоянието 2 и последното действие е червена стрелка, тогава може да сме дошли от състоянията 1 и 5. Ако знаем, че в състоянието 1 агента никога не би избрал червена стрелка, тогава предишното състояние трябва е било 5.

При event-driven моделите ние не правим разлика между миналото и бъдещето. Събираме статистика и за едното и за другото, опитваме се да предскажем и едното и другото.

Много по-лесно е да опишем света и агента като единна система, отколкото да опишем само света без агента. Например, вие няма да си боядисате косата зелена и това е едно просто описание на вас и на света, в който вие живеете. Дали няма да го направите защото сте достатъчно разумен или не можете да го направите защото нямате зелена боя, това е без значение. Ако някой ви попита дали утре ще сте със зелена коса, можете спокойно да му кажете, че това не може да се случи. Дали това го знаете от анализа на вашето собствено поведение или от някакви ограничения, които са наложени от света, това е без значение. Важното е, че косата ви не може утре да е зелена.

Неформално описание

Първо ще опишем задачата неформално. Имаме един свят (environment) и един агент, който живее в този свят. Света е един ориентиран граф подобен на този на фигура 5. Агентът се движи от състояние на състояние по стрелките. Състоянията и стрелките са етикирани с някакви етикети. На фигура 5 вместо етикети сме използвали различни цветове. Докато се движи агента въвежда (observe) информация (това е етикета на състоянието, в което се намира) и действа (избира етикета на следващата стрелка, по която ще мине). Да отбележим, че агента избира само етикета (цвета) на стрелката, но не и конкретната стрелка. Например в състояние 6, ако избере да продължи със синя стрелка, то не се знае дали ще отиде в състояние 5 или в 7. Може някоя възможност да не съществува. Например от състояние 2 не излиза нито една червена стелка. Тоест, в това състояние агента няма право да избере да продължи по червена стрелка и задължително трябва да избере синя.

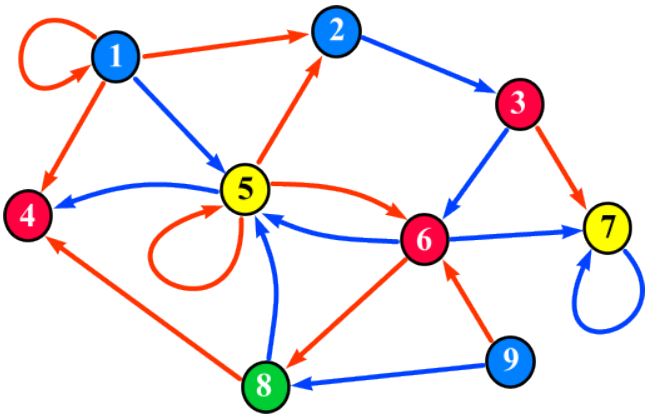

Figure 5

Пътя на агента в ориентирания граф ще наречем живота на агента. Всяко състояние може да бъде начало или край на живота, но има състояния, които, ако участват в пътя, задължително са неговото начало. Такива състояния ще наречем абсолютно начало. Това са състояния, в които не влизат стрелки. Тоест, преди тях не може да има минало. На фигура 5 такова състояние е 9. Състоянието 1 не е абсолютно начало, заради червената примка (loop). Аналогично, състоянията, от които не излизат стрелки, ще наречем внезапна смърт. Такова състояние, ако участва в пътя, то задължително е последното. След

такова състояние няма бъдеще. На фигура 5 такова състояние е 4. Състоянието 7 не е внезапна смърт, заради синята примка.

Идеята да наречем състоянията, от които не излизат стрелки, „внезапна смърт“ е това, че живота на агента може да завърши по два начина – когато ние го изключим или когато той попадне в такова състояние и спре. Когато ние изключваме агента, ще казваме, че имаме естествена смърт, а когато той сам спре, ще казваме, че това е внезапна смърт. (В естествения език, когато някой ни изключи не говорим за естествена смърт, а за убийство. Тук, обаче, ще предполагаме, че изключваме агента, когато е минало много време. Тоест, изключването ще го асоцииираме с умиране от старост и затова ще го наричаме естествена смърт.)

Нека отбележим, че живота не е задължително да започва от някое абсолютно начало. Миналото може да е безкрайно, по същия начин както безкрайно може да бъде и бъдещето. Тоест, живота може да е един безкраен път без начало и без край.

При MDP и при RL разглеждаме два случая – когато агента вижда всичко (Full Observability) и когато той вижда частично (Partial Observability). На фигура 5 бихме имали Full Observability, ако всички състояния имаха различни цветове или ако агента виждаше не цвета, а номера на състоянието (тоест, ако агента виждаше кое точно е състоянието, в което се намира).

В повечето статии се предполага, че при MDP имаме частния случай (special case) на Full Observability, а когато имаме по-общия случай (generalization) на Partial Observability те използват термина partially observable Markov decision process (POMDP). В тази статия ще е обратното. С MDP ще означаваме по-общия случай, а когато имаме Full Observability ще използваме термина fully observable Markov decision process (FO MDP).

Аналогично, тук ще предполагаме, че при Reinforcement Learning имаме Partial Observability. Когато разглеждаме случая на Full Observability, специално ще отбелязваме това.

Начално състояние

При Reinforcement Learning (RL) се говори за начално и за текущо състояние. За да опишем един път (живот) на нас ни е нужно състояние, от което да тръгнем. Обикновено авторите предпочитат началното състояние, защото предполагат, че напред към бъдещето е по-лесно да се предсказа, отколкото назад към миналото. Тоест, да се върви по стрелките е по-лесно, отколкото да се върви срещу стрелките. Например, при Markov decision process (MDP) имаме вероятности, които ни казват кое очакваме да е следващото състояние, ако вървим напред по определено действие, но нямаме вероятности в обратната посока. В тази статия няма да говорим за начално, а само за текущото състояние. Това ще е така, защото ще предполагаме, че миналото и бъдещето са равноправни и се предсказват еднакво трудно.

Причините поради които няма да се интересуваме от началното състояние са много. Начално състояние може въобще да нямаме, защото миналото може да е безкрайно. Кое е началното състояние на човека? Дали това е момента на раждането или момента на

зачеването? В модела на света, който използваме има моменти, които са се случили преди нашето раждане. Тоест, в нашия модел нашето раждане не е абсолютно начало. Има един модел на света, в който има абсолютно начало и това е теорията на Големия взрив. Дали тази теория е вярна? Ние няма да си задаваме въпроса вярна ли е една теория, а дали тази теория ни върши работа (дали описва света). Има много различни теории, които описват света. За всяка от тези теории ще считаме, че е вярна, стига да не може да се направи експеримент, който да противоречи на теорията.

Тоест, при RL ще се питаме какъв е света и къде сме в момента. Обикновено си задаваме въпроса „къде съм“, а не къде съм бил, когато съм се родил. Тоест, ще търсим текущото състояние, а не началното.

Какво е история

Какво е живот? Това е цялата история, тоест история, която не може да бъде продължена, защото е безкрайна или защото няма следваща стъпка – например защото историята е стигнала до внезапна смърт или защото сме изключили агента (естествена смърт).

Какво е история? **Дефиниция:** Съкратена история ще наричаме редицата от действия и наблюдения от началния момент до текущия, заедно с последното коректно действие.

$$a_1, v_1, a_2, v_2, \dots, a_t, v_t, a_{t+1}$$

Дефиниция: История ще бъде когато добавим и некоректните ходове, които сме пробвали и за които знаем, че са некоректни.

$$bad_0, a_1, v_1, bad_1, a_2, v_2, \dots, bad_{t-1}, a_t, v_t, bad_t, a_{t+1}$$

В тази редица a е действие, а bad е множество от действия. Нека отбележим, че следващата стъпка започва с поредния коректен ход. Множеството bad може да се приеме като част от наблюденията на предишната стъпка.

Дефиниция: Пълна история ще бъде когато множествата на некоректните ходове, които сме пробвали ги заменим с множествата на всичките некоректни ходове.

$$full_0, a_1, v_1, full_1, a_2, v_2, \dots, full_{t-1}, a_t, v_t, full_t, a_{t+1}$$

Каква е разликата между история и пълна история? Първото е това, което сме видели по пътя, а второто е това което бихме могли да видим, ако гледахме по-внимателно. Имаме $bad \subseteq full$. Тоест, ходовете, които сме пробвали и за които знаем, че са некоректни са част от всички некоректни ходове. Бихме могли да пробваме още ходове, така че да стане $bad = full$, но така може без да искаме да изиграем някой коректен ход и историята да стане съвсем различна.

Дефиниция: Локална история с дължина k ще наричаме последните k стъпки на някоя история. Тоест, края на някаква история.

Дефиниция: Приблизителна история. Това ще бъде някакво непълно описание на историята. Предполага се, че историята е твърде дълга и ще е трудно да я помним цялата. Затова ние помним само края на историята или помним само някакви по-важни събития (кога са се случили) или помним някаква статистика за историята. Много често на нас цялата история не ни е нужна и приблизителната история ни е достатъчна, за да направим модел на света и да планираме бъдещите си действия.

Нека отбележим, че за агента не е важно през кои състояния е преминал, а какво е видял. Макар живота да отговаря на път в ориентирания граф (модела на света), на два различни пътя може да отговарят на един и същи живот.

Целта на агента

Каква е целта на агента? Целта е по-добър живот.

Кой живот е по-добър? Трябва ни една релация, която да сравнява животите. Тази релация трябва да е квази-ред (preorder or quasiorder). Тоест тя трябва да е рефлексивна и транзитивна (reflexive and transitive). Ако добавим и антисиметричност (antisymmetry) ще получим частична наредба (partial order). Антисиметричността не ни е нужна, защото няма да е проблем два различни живота да са еднакво успешни.

Бихме искали тази релация да притежава някаква монотонност. Нека разделим живота на две половини и нека вземем два живота. Ако първата половина на първия живот е по-добра от първата половина на втория и същото за вторите половини, то бихме искали първият живот да е по-добър от втория. Обратното би било доста странно.

Подобна релация можем да получим на базата на събирането на награди (rewards) и наказания (regrets). В тази статия няма да казваме награди и наказания, а ще казваме оценки. Наградите са положителни оценки, а наказанията са отрицателни оценки. Обаче има оценки, за които не можем да кажем дали са положителни или отрицателни. Всичко е относително. Например ако получите тройка в училище, дали това е положителна или отрицателна оценка. За едни е положителна, а за други е отрицателна.

Така, събирайки оценките можем да получим едно число, което да е оценка на целия живот. В [3] това число се получава точно като сумата от оценките, но там дължината на живота е фиксирана и сумата е еквивалентна на средното аритметично. В [4] това число се получава като средното аритметично от оценките. В по-късни статии се използва discount factor, което е грешка. По-надолу ще обсъдим защо това е грешка.

В повечето статии се предполага, че на всяка стъпка агента получава оценка. Това е така в [3, 4] и при много други статии. Подобно предположение може да се приеме като прийом за опростяване на изложението, но по принцип идеята да получаваме оценка на всяка стъпка е нелогична и неестествена. Например, в училище не ни изпитват всеки ден.

Можем да приемем, че в случаите, когато нямаме оценка, оценката е нула, но тогава всичките тези нули съществено ще променят средното аритметично. Това би променило и стратегията ни. Например, ако играем шах и ремито е нула, тогава ще е все едно дали сме постигнали реми или играта продължава. Ще има и голямо значение колко дълга е

партията. Тогава, ако средната ни оценка е отрицателна ще се опитваме да удължим партията максимално. Ако средната оценка е положителна, тогава ще се стремим партиите ни да са колкото е възможно по-къси. Естествено е ние да се опитваме да приключим партията възможно най-бързо, но в този случай ще сме склонни да рискуваме победата само и само за да приключим по-бързо. На шахматните турнири се изчислява средното аритметично от победите, загубите и ремитата без да се отчита колко хода е била средната партия.

Тоест, повечето статии предполагат, че оценката е реално число, а ние ще предполагаме че тя е реално число или константата Undef. Тоест, ще предполагаме, че може да имаме, а може и да нямаме оценка.

Различни критерии

В повечето статии се предполага, че имаме един единствен критерии, по който оценяваме живота. Например при икономическите задачи критерият са парите и най-добра е тази стратегия, при която печалбата е най-голяма. Нека да предположим, че имаме два критерия. Например искаме най-добра печалба, но без да влизаме в затвора.

Ще разглеждаме цели, които се определят от повече от един критерии. Например, ако имаме двама души и единия има повече пари, а другият има повече деца, то кой е по-успешен? Ако двата критерия са равноправни, то за да бъде един живот по-добър от друг, то той трябва да е по-добър и по двата критерия. Така релацията „по-добър живот“ става нелинейна, но ние казахме, че искаме тази релация да е квази-ред и не е задължително тя да е линейна.

Може да имаме два критерия и единия да има приоритет пред другия. Например, ако пишем програма за кола, която сама ще се движи, то целта ще е по-малко закъснения и по-малко убити хора. Можем да подходим като оценим закъсненията и човешкия живот с пари и да търсим стратегията с минимален разход. Тоест, може да сведем двата критерия към един. Когато ние караме кола, точно така постъпваме. Оценяваме стойността на закъснението и когато бързаме увеличаваме риска. Тоест, склони сме да поемем по-голям риск да убием някого, когато закъсняваме. Въпреки, че така постъпват хората шофьори, вашата програма не може да постъпи по същия начин. Ако оцените човешкия живот на някаква сума пари, то ще ви обвинят в безчовечност. Затова трябва да дадете приоритет на втория критерии и да сравнявате две стратегии на базата на броя на жертвите. При равен брой жертви, тогава по-добрата стратегия ще е тази с по-малките закъснения. Тоест, ние ще оценим човешкия живот като безкрайно по-ценен от закъснението. По този начин получаваме линейна наредба, но с два критерия.

Тоест, ако имаме n критерия, то оценката трябва да е n-мерен вектор. Всяка от координатите трябва да има стойност реално число или константата Undef. Накрая дали живота е по-добър се определя, като се изчисли средното аритметично на всяка от координатите и така получения вектор описва резултата от целия живот.

В тази статия, за да опростим изложението ще предполагаме, че критерия е само един. Все пак, ще отбележим, че Sutton в [2] не е прав като изказва “the reward hypothesis”: all goals

and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a single externally received number (reward).

Discount factor

Казахме, че използването на discount factor при дефиницията на MDP и на RL е грешка. Това е често срещаната грешка когато бъркаме въпросите „какво“ и „как“. Какво искаме да направим и как ще го направим? Например, къде искаме да отидем на почивка? Искаме да отидем в Хонолулу. После си казваме, че Хонолулу е твърде далеч и решаваме, че искаме да отидем някъде по-наблизо. Това е грешка, защото бъркаме въпросите „какво“ и „как“. Какво искаме и как ще го постигнем. Ние искаме да отидем в Хонолулу, а това че ще отидем някъде по-наблизо не променя това къде искаме да отидем.

Discount factor произлиза от една съвсем естествена стратегия, която ни казва, че по-близките награди са по-важни от по-далечните. Дори има народна поговорка, която изразява тази стратегия: „Не изпускай питомното, за да гониш дивото!“.

Discount factor е свързан със стратегията, а не със смисъла на живота. От кой момент нататък да приложим discount factor? Това е текущия момент, но текущия момент не е фиксиран, а се променя. Ако приложим discount factor от началния момент, то началото на живота би било много по-важно от края му, което не е логично.

Каква е била стратегията на Наполеон непосредствено преди битката при Ватерло. Стратегията е била: „Да спечелим битката при Ватерло, а после ще видим.“ Тоест, непосредствено преди битката, тя е била най-важното нещо за него, но ако оценяваме целия живот на Наполеон, то тази битка няма да има чак толкова голяма тежест.

Тоест, да се използва discount factor, за да се определи кой живот е по-добър е грешка. Най-малкото, защото не се знае кой да е момента, от който да приложим този коефициент и какъв да е размера му. Например когато сме по-неуверени ние бихме избрали по-малък discount factor и бихме се стремили към най-близките награди. Колкото по уверени ставаме, толкова по-надалеч в бъдещето ще гледаме и ще сме по-склонни да пуснем питомното, за да гоним дивото. Тоест, колкото сме по-уверени, токова по-близък до единицата ще е коефициента на обезценка, който ще изберем.

Хубавото на discount factor е, че когато го използваме можем да оценим безкрайния живот. Получаваме едно число, което е сумата на една геометрична прогресия. Как да сравним дава безкрайни живота и да кажем кой от тях е по-добър? Това ще стане по следния начин:

$$Live1 \geq Live2 \Leftrightarrow \exists n \forall (k \geq n)\ begin(Live1, k) \geq begin(Live2, k)$$

Тук *begin(Live, k)* е началото на *Live*, което е с дължина *k*. Тук релацията $\geq$ е нашата квази-ред релация „по-добър“, която ни казва кой живот е по-добър. Предполагаме, че сме дефинирали тази релация за крайните животи и я продължаваме и за безкрайните. Горната формула може да се използва и за сравнение между краен и безкраен живот, ако приемем че когато *k* е по-голямо от дължината на живота, тогава *begin(Live, k)= Live*.

Какъв е модела?

Казахме, че модела на света ще бъде ориентиран граф. Изникват няколко въпроса. Първо дали модела на света да е детерминиран ли недетерминиран. Второ, ако сме предпочели недетерминираните графи, то каква да бъде случайността. В [5] обсъдихме това, че имаме два вида случайности. Първия е, когато знаем точната вероятност за избора на всяка от стрелките (MDP). Втория е, когато знаем кои са възможните стрелки, но не знаем с каква вероятност всяка от тези стрелки може да бъде избрана. На фигура 5 е изобразен втория случай, защото не са дадени вероятности по стрелките с еднакъв цвят. В [5] се разглежда и комбинацията от тези две вероятности, когато не знаем точната вероятност, но имаме интервал и знаем, че вероятността е в този интервал.

В [5] доказахме, че при RL детерминираният модел и моделите с различните видове вероятност ни дават еквивалентни дефиниции. Тоест, не можем да познаем дали света е детерминиран или недетерминиран. Разбира се, това е при условие, че живеем в света само един живот. Ако живеем в света два живота и ако двата пъти правим едно и също, то в детерминирания свят ще се получи същия живот, докато в недетерминирания, почти сигурно, ще се получи различен живот.

Важно предположение е, че при RL живеем само един живот и имаме само една история, на базата, на която трябва да правим изводи. Ако допуснем, че сме живели няколко пъти в един и същи свят и че имаме няколко истории на базата на които да трупаме опит, то се получава съвсем различна задача, при която детерминираният и недетерминираният модел не са еквивалентни.

В тази статия ще започнем с детерминиран модел, който ще наречем „съвършен модел“. След това ще дефинираме недетерминиран модел, който ще наречем „модел със случайност“. После ще заменим действията със събития и ще получим event-driven модел.

Къде е оценката? В тази статия оценката ще е част от наблюдението. В повечето статии това не е така. Обикновено имаме две различни наблюдения, първото от които се казва наблюдение и то определя в кое състояние сме (Full Observability). Второто наблюдение се казва оценка. Първото наблюдение е етикет към състоянието, а второто е етикет към стрелката. В тази статия ще имаме едно наблюдение и оценката ще е част от него.

Некоректни ходове

Дали модела да допуска некоректни ходове или да предполагаме, че всеки ход е коректен? За всеки модел с некоректни ходове можем да направим еквивалентен на него модел, в който всички ходове да са коректни. В [6] направихме нещо подобно като построихме тоталния модел. Тоест, добавихме едно наблюдение *bad*, което да се получава винаги след некоректен ход.

Нека за агента да имаме две програми – тотално AI и частично (partial) AI. Първата програма да изисква всички ходове да са коректни, а втората да допуска некоректни ходове. Ако дадем на тоталното AI да се бори със света, който е представен като тотален модел, то така бихме направили задачата на агента доста по-трудна. Той ще търси зависимости в реда на пробваните некоректни ходове. Няма да намери такава зависимост,

защото нея я няма, но няма да престава да я търси. Ще пробва да повтори два пъти един и същи некоректен ход. Накрая ще се научи, че от това нищо ново не се случва и ще спре тези опити. Този агент няма да знае, че пробвайки един некоректен ход той не променя нищо (освен, че разбира че хода е некоректен, но ако това вече го знае, то той не променя нищо). Тези неща също могат да се научат, но винаги когато този агент е изпаднал в безизходица и се чуди какво да прави ще пробва пак разни неща, които е безсмислено да се пробват. Живота на частичното AI в частичния модел ще е много по-лек, защото той по рождение ще знае някои неща за некоректните ходове и няма да има нужда сам да открива и да учи тези неща.

Затова е по-добре да допуснем, че имаме некоректни ходове. Това прави света много по-прост и по-лесно разбираем. Спестяваме на агента търсенето на зависимости, които не съществуват и тяхното търсене би било просто загуба на време. Освен това некоректните ходове са много хубав пример на полувидимо събитие (за тях ще стане дума по-надолу). Също така некоректните ходове са хубав пример за тестови състояния (за тях се говори в [5], но ще говорим още за тях в следващата статия).

Reinforcement Learning

Нека да кажем, коя ще е формалната дефиниция на Reinforcement Learning, която ще използваме. Ще кажем какво ни е дадено и какво търсим. Дадени са ни:

A – множество от възможните действия на агента.
V – множество от възможните наблюдения.
$Reward : V \to \mathbb{R} \cup \{Undef\}$ (това е функция, която за всяко наблюдение ни дава оценка или *Undef*)
H – история или приблизителна история на случилото се до момента t.
Дадено ни е още, че съществува съвършен модел M на света и този модел се намира в някакво текущо състояние s_t.

Какво търсим? Трябва да отговорим на три въпроса:
1. Какъв е света? (трябва да намерим модела на света)
2. Къде съм? (трябва да определим текущото състояние на света)
3. Какво да правя? (трябва да решим какво ще е нашето следващо действие, както и по-нататъшните ни действия, като целта е да постигнем максимална стойност на средноаритметичното на оценките)

Забележка: В тази статия се занимаваме основно с първите два въпроса, а оценката е свързана само с отговора на третия въпрос. Затова по-надолу няма да става дума за оценки.

Съвършен модел

Нека да кажем какво представлява съвършеният модел на света:
S – множеството от вътрешните състояния на света.
s_t – текущото състояние на света.
$G = <S, R>$ – тотален и детерминиран ориентиран граф.
$R \subseteq S \times A \times S$

View: $S \to V$
Incorrect: $S \to P(A)$

На всяко ребро (стрелка) има етикет, който ни казва кое е действието, което трябва да извършим, за да преминем по тази стрелка. Стрелките и техните етикети се определят от релацията R. На всеки връх (състояние) съответстват два етикета, които казват какво виждаме и кои ходове са некоректни в това състояние. Тоест, етикетите са наблюдението в това състояние и множеството от невъзможни действия. Тези етикети на състоянието се определят от функциите *View* и *Incorrect*.

Ще предполагаме, че ориентирания граф G е тотален и детерминиран. Тоест, от всяко състояние, по всяко действие има стрелка, която излиза от това състояние и тази стрелка е единствена. Тук стрелките, които отговарят на некоректни ходове са в известна степен излишни, защото никога няма да ги използваме, но предполагаме, че и те съществуват, защото по-долу ще разгледаме друг модел на света, в който множеството некоректните ходове няма да е постоянно, а ще се променя. Тоест, при едно и също състояние в различни моменти ще можем да имаме различни некоректни ходове.

Нека отбележим, че когато имаме съвършен модел на света и когато знаем кое е текущото състояние на света, тогава бъдещето е напълно определено. Тоест, ако знаем кои ще бъдат действията на агента, мажем да кажем точно какво ще се случи. За разлика от бъдещето, миналото не е напълно определено, защото може да имаме много различни начални състояния, които след историята H да водят до това текущо състояние. Дори и пълната история не е определена, защото може различни начални състояния да имат различна пълна история. Разбира се, различните пълни истории трябва да са съгласувани с историята H (тоест трябва да удовлетворяват условието $bad \subseteq full$). Ще предполагаме, че имаме поне едно възможно начално състояние, което след историята H да води до текущото състояние, защото предполагаме, че тази история все пак се е случила в този модел.

Ако освен текущото знаем и началното състояние, тогава от съвършения модел можем да получим така наречения съкратен модел. Това ще е модела, в който сме изхвърлили всички състояния и стрелки през които не сме минали прочитайки историята H. В съкратения модел можем да направим статистика като преброим колко пъти сме минали по всяка една от стрелките. Чрез тази статистика ние ще можем да предскажем миналото, бъдещето и поведението на агента. Когато знаем в определено състояние колко пъти агента е избрал едно действие и колко пъти друго действие, ние може да предвидим какво би направил агента, ако пак попадне в това състояние. Тоест, ние ще се опитвам да предскажем собственото си поведение, защото както казахме, ще разглеждаме света и агента като единна система.

Съкратения модел може да не е съвършен. Може да попаднем на стрелка, по която никога не сме минавали (тоест, такава която я няма в съкратения модел). Въпреки всичко, съкратения модел е това, което бихме могли да знаем, защото ако по една стрелка никога не сме минавали няма откъде да знаем накъде води тя.

Случайни величини

Нашето желание да намерим съвършен модел на света е една твърде амбициозна задача, която би била възможна само, ако света е изключително прост. Нас, естествено, ни интересуват по-сложни светове и затова е разумно да предположим, че света е толкова сложен, че ние никога няма да можем напълно да го разберем.

Нека допуснем, че има предел на нашето познание и има неща, които ние никога няма да можем да предскажем. Тези неща ние ще наречем случайни величини (random variables). Ще предположим, че това са зависимости, които са толкова сложни, че няма как да ги отгатнем. Например, ако хвърля зар, то предполагам, че ще се падне число от 1 до 6 с вероятност 1/6. Ако имах съвършен модел, то бих могъл да кажа точно колко ще се падне. Има такъв съвършен модел и след като хвърля зара ще знам кой е той, но на мен ми трябва да го знам още преди да съм хвърлил зара. Няма как преди да хвърля зара да знам какво ще се падне.

Трябва да отбележа, че ако аз бях много много умен, щях да мога да кажа какво ще се падне. Щом не мога да кажа, значи или не съм достатъчно умен, или нямам достатъчно информация на базата, на която да кажа какъв ще е зара.

Допускам, че не съм достатъчно умен и най-доброто предположение, което мога да направя е число от 1 до 6 с вероятност 1/6. Дори и това да е най-доброто предположение, което аз мога да направя, това не значи че аз няма да търся и по-добро. Може да предположа, че зара е крив и дава 6 с по-голяма вероятност от 1/6 или че ако духна върху зара преди да го хвърля, то вероятността да се падне шестица се увеличава. Последното предположение се приема за суеверие (тоест за невярно), но тук ще търсим статистическа връзка между събития и ако статистиката показва, че духването върху зара ни помага, то това се приема за доказателство, дори да се окаже, че това се е случило случайно.

Какво е случайната величина? Да предположим, че аз сега ще ви кажа черно или бяло. Какво очаквате да чуете – черно или бяло? Предполагам, че идея си нямате какво ще ви кажа. Тест, вие очаквате да чуете бяло с вероятност в интервала [0, 1].

Ако знаете, че аз имам един зар с една бяла страна и пет черни. Знаете, че аз ще хвърля този зар и така ще определя дали да ви кажа черно или бяло. Тогава вие ще очаквате бяло с вероятност 1/6. Да предположим, че аз имам втори зар с две бели и четири черни стени. Ще хвърля един от тези два зара и така ще определя дали да ви кажа черно или бяло. Сега предполагате да чуете бяло с вероятност в интервала [1/6, 1/3]. Да предположим, че знаете че е по-вероятно да хвърля първия зар, отколкото втория. Сега очаквате вероятност в интервала [1/6, 1/4).

Да предположим, че аз казвам веднъж бяло, веднъж черно. Тоест вие ще очаквате да чуете обратното на това, което съм казал последния път. Това обаче е зависимост с памет. Аз бих искал да опиша случайните величини без параметри. Случайната величина с параметри е функция, която при дадени параметри ни връща случайна величина без параметри. По-надолу ще опишем един оракул α, който зависи от миналото, бъдещето и от текущото състояние. Тоест, миналото е един от възможните параметри.

Да предположим, че аз когато съм в добро настроение казвам бяло или поне вероятността да кажа бяло се увеличава. Това добавя към модела още един параметър, който ще бъде моето настроение. Тоест, можете да се опитвате да предскажете какво е моето настроение или ако моето настроение е напълно непредвидимо, тогава имате случайна величина, която не използва моето настроение като параметър. Тази величина ще зависи от моето настроение, но щом то е непредвидимо, то няма да е параметър. Ще бъде скрит параметър или параметър, който няма да се опитваме да предскажем.

Може да предположите, че това което ще ви кажа е възможно най-лошото. Подобно предположение се използва често. Например в програмите играещи шах в алгоритъма Min-Max се предполага, че противника ще изиграе възможно най-нежелания от нас ход. Тоест, предполагаме, че имаме противник, който ни мисли лошото. Не винаги предполагаме най-лошото. Понякога предполагаме най-доброто. Например, ако се прибирате вкъщи след дълъг път, какво очаквате да има за вечеря? Очаквате най-доброто, защото там има някой, който ви обича и който ви е приготвил това, което вие най-много обичате.

Да предположим, че очаквате най-доброто. Ако бялото по някаква причина е по-добро от черното, то вие бихте очаквали да ви кажа „бяло“. Може да не знаете кое е по-доброто и това да се разбере в последствие. Тогава вашето очакване зависи от бъдещето. Тоест, зависи от друг параметър.

Случайна величина без параметри

Въпросът е какво е случайна величина без параметри (RVWP). Повечето автори предполагат, че RVWP си има точно определена вероятност (или разпределение, ако възможните стойности са повече от две). Такова е и предположението при MDP. Там случайно се избира следващото състояние, но тази случайност не е съвсем случайна, защото се предполага, че вероятността да се избере определено състояние е точно определена.

Тук ще предполагаме, че имаме две нива на неопределеност. Първото ниво е, когато не знаем какво ще се случи, но знаем точната вероятност p, с която това ще се случи. Второто ниво на неопределеност е, когато не знаем дори и вероятността p. Може такава вероятност въобще да няма, а може да има, но ние да няма откъде да я знаем.

Когато предвиждаме едно случайно събитие, да кажем че то си има някаква вероятност p, с която ще се случи. Ако сме достатъчно умни и имаме нужните данни, ние ще предскажем тази вероятност. До, но събитието има и конкретна стойност, която ще се случи и ако ние сме много много умни ще можем да познаем дори и тази конкретна стойност. С първото ниво на неопределеност ние приемаме, че не можем да познаем какъв ще е резултата, а с второто ниво на неопределеност приемаме, че дори и вероятността не можем да познаем.

Може ли едно събитие въобще да няма точна вероятност? Ако има някаква вероятност p и ако наблюдаваме събитието безкрайно дълго, то статистически получената вероятност трябва да клони към p (закон за големите числа). Ако наблюдаваме събитието безкрайно дълго статистиката може да не клони към определена стойност (ще имаме limit inferior и limit superior, но те може да не са равни).

Много лесно е да си представим ситуация, в която не знаем точната вероятност, а само знаем, че тя е в интервала *[0, 1]* или *[a, b]*. Много трудно е да конструираме физически експеримент, който ни дава такава вероятност и тази вероятност да е най-доброто предположение, което можем да направим. Например нека вземе една редица от нули и единици, за която limit inferior и limit superior от средното аритметично са съответно 0 и 1. Тази редица вероятно ще се движи зигзагообразно и когато вероятността за последните 100 е голяма ще очакваме следващото число да е единица също с голяма вероятност. Тоест, тази редица ще има вероятност в интервала *[0, 1]*, но това няма да е най-доброто описание на случайната величина, защото ще имаме и по-добро, което ще получим от последните 100 резултата.

Ако искаме да направим физически експеримент, който ни дава точно определена вероятност *p*, то това е лесно. Просто конструираме зар, който ни дава вероятност *p* и готово. Не само, че това ще е случайната величина описваща този експеримент, но това още е най-доброто възможно описание на експеримента.

Как да конструираме физически експеримент, който ни дава вероятност в интервала *[0, 1]* и да няма по-добро описание на този експеримент? Ще предположим, че срещу агента стои едно същество, което казва 0 или 1, но не случайно, а така че да обърка агента и той да не може да познае дали следващото ще е нула или едно, и той дори да не може да каже с каква вероятност очаква следващото да е едно. Ще предполагаме, че съществото е много по-умно от агента и действително успява да го обърка.

Това, което направихме, не е коректна математическа дефиниция, защото колкото и да си умен винаги има някой, който е по-умен от теб. Тоест, съществото зависи от агента. Ако съществото може да прочете мислите на агента и да познае, че той очаква единица в интервала $[a, b] \subset [0, 1]$, тогава може да го обърка като играе така, че да излезе, че вероятността е извън този интервал. Да прочете мислите звучи екзотично, но ако агента е алгоритъм (програма), тогава съществото може да изпълни тази програма и да разбере какво агента очаква да се случи.

Така направихме модел на събитие, чиято вероятност е в интервала *[0, 1]*. Ако искаме да направим събитие, чиято вероятност да е в интервала *[a, b]*, то ще конструираме два зара с вероятности *a* и *b* и ще дадем на съществото тези зарове. То ще избира, кой от заровете да хвърли, като целта му пак ще е да обърка агента.

Другите автори предполагат, че случайната величина без параметри има точна вероятност (разпределение), а ние ще предполагаме, че има RVWP точно определен интервал на вероятността (или разпределение от интервали). Разбира се, тези интервали от вероятности ще трябва да изпълняват неравенствата описани в [5]. Ще предполагаме, че RVWP не зависи нито от миналото нито от бъдещото. Както е хвърлянето на зар, то не зависи нито от хвърлянията преди, нито от хвърлянията след това.

Модел със случайност

Предположихме, че има случайност, която не може да бъде предсказана. Сега ще направим втори модел на света, който включва тази случайност. Първо, ориентираният граф вече няма да е детерминиран (в дефиницията на MDP графа също не е детерминиран). Второ

това, което наблюдаваме в състоянието s, вече няма да е константа функция, а ще е случайна величина (при POMDP наблюдението също е случайна величина). Трето, множеството на некоректните ходове в състоянието s, също ще бъде случайна величина. Тези три случайни неща ще се определят от три оракула (α, β и χ). Оракулите ще са случайни величини с параметри, а при конкретна стойност на параметрите те ще са случайни величини без параметри (RVWP).

Оракула α казва при дадено състояние s_t и дадено действие a_{t+1} какво ще е следващото състояние. Това до голяма степен е определено от графа G. Оракула α е задължен да избере една от възможните стрелки. Така че избор за оракула има само когато прехода е недетерминиран.

Оракула β казва при дадено състояние s_{t+1} какво ще е наблюдението в момента $t+1$. Тук сме написали $t+1$ вместо t, защото трите оракула използват едно и също *Past* (a_{t+1} е края на това *Past*.) Първо оракула α казва кое ще е следващото състояние и после оракулите β и χ казват какво ще се види в това състояние и кои ще са некоректните ходове.

Оракула χ казва дали събитието e се е случило в момента t. Тук ще използваме χ за събития от типа дали определен ход е некоректен. По-надолу ще има и други събития, които също ще се определят от χ.

Ето дефиницията на модел със случайност:

S – множество от вътрешните състояния на света.
s_t – текущото състояние на света.
$G = <S, R>$ – тотален ориентиран граф (недетерминиран).
$R \subseteq S \times A \times S$
$\alpha(Past, s_t, a_{t+1}, Future) \rightarrow s_{t+1}$
$\beta(Past, s_{t+1}) \rightarrow v_{t+1}$
$\chi(Past, s_{t+1}, e) \rightarrow \{true, false\}$

Идеята на оракулите е, че агента не знае в кое състояние е света, какво ще види в това състояние и кои ходове ще са коректни. Агента не знае, но света знае всичко. Тоест, света разполага с тези оракули и може да каже какво ще се случи. Да предположим, че света има на разположение един съвършен модел на света. Такъв модел би определил напълно трите оракула.

Забележка: Тук за всяко събитие e оракула χ дава отделна случайна величина. Дефиницията остава впечатлението, че това са независими случайни величини, а това може и да не е така. Две събития може да са свързани. Например събитието a може да се случва само когато се е случило събитието b. Друг пример е когато a и b никога не се случват едновременно. Наблюдението също може да има връзка със събитията. Може би е добре дефиницията да се усложни и оракулите β и χ да се обединят в един по-сложен оракул, но това няма да го правим.

Оракулите зависят от миналото. Ще предполагаме, че *Past* е пълната история. (Предполагаме, че използваме пълната, а не обикновената история, защото оракулите зависят от това какво се е случило, а не от това което агента е видял. Агента знае само

обикновената история, а света знае пълната история. Предполагаме, че агента ще се научи да познава кои ходове са коректни и той на практика също ще знае пълната история.)

Ще предполагаме, че оракула α зависи и от бъдещето. Това звучи доста екзотично, но ако съобразим кое е причината и кое е следствието, това че оракула вижда бъдещето не е чак толкова странно. Да вземем примера с пощальона и скитника. Странно би било, ако кажем, че оракула е погледнал в бъдещето, видял е че в бъдеще ще се появи доказателство за това, че пощальона е убиеца и затова оракула е решил убийството да бъде извършено от пощальона. По-логично би звучало, ако кажем, че убийството е извършено от пощальона и затова в бъдещето се е появило доказателство, че той е убиеца.

Това, че оракула зависи от бъдещето ще повлияе на отпечатъка на модела. Видяхте във фигура 2, че това че оракула избира стрелките по този начин осигурява това, че на следващата стъпка от състоянието 1 не излиза стрелка по събитието *b*. Тоест, това събитие никога не се е случило, когато сме били в състояние 1. Ако оракула не се държеше по този начин, то от състоянието 1 щеше да излиза стрелка по *b*.

Ще предполагаме, че на оракула α е дадено цялото бъдеще (до края на живота или до безкрайност). Въпреки това, ние нямаме цялото бъдеще, а само бъдещето до текущия момент *t*. В условието на задачата ни е дадена историята до момента *t*. Ако искаме да разберем какво ще каже оракула α в момента *t-k*, тогава ще разделим историята на минало от 0 до *t-k* и бъдеще от *t-k* до *t*. Трябва да отбележим, че решението на оракула зависи обикновено само от близкото бъдеще. В примера с убийството, ако не се докаже кой е убиеца скоро след убийството, вероятно въобще няма да се докаже. Веднъж докаже ли се нещо, после няма как да се докаже обратното.

Ако използваме това, че живота се живее само веднъж може да определим оракула β като функция. Разбира се, тази функция ще е дефинирана само за историите, които са се случили в този живот. За останалите истории ще трябва да си измислим стойностите на тази функция (да ги дефинираме както и да е). Ако знаем пълната история на живота, то ще можем да определим и оракула χ. Ако имаме само обикновената история, то този оракул ще можем да го определим само частично. Оракула α можем да го определим по произволен начин. Можем да вземем произволен граф *G* и да дефинираме оракула α по този граф. Това би определило α винаги освен в случаите на недетерминирано разклонение (тогава ще можем да определим оракула както си поискаме). Пример за произволен граф е, когато имаме само едно вътрешно състояние. Дори и този граф е един възможен модел на света (фигура 1 е такъв модел).

Трябва да отбележим, че ако имаме един и същи ориентиран граф, но различен оракул α, то модела вероятно ще е различен, защото отпечатъка на модела може да е различен. Ние няма да търсим различни оракули α, а ще търсим различни отпечатъци. Ако два оракула дават еднакъв отпечатък на модела ще ги приемаме за еднакви. Обратното, ако ни е даден отпечатъка по него може да построим оракул, който индуцира този отпечатък (ако въобще има такъв оракул, разбира се). Това ще стане по следния начин. От възможните избори на оракула ще махнем тези, които са невъзможни при конкретния отпечатък и при бъдещето, което ни е дадено. Ако са останали още възможни избори, ще изберем един от тях. Ще изберем този един с вероятност съответстваща на вероятността този избор да е възможен при този отпечатък и при това бъдеще.

Отпечатък

Нашата идея е, че ориентирания граф G по някакъв начин описва света. Тук се оказа, че произволен граф G може да бъде модел на света! Истината е, че всеки граф може да е модел, но далеч не всеки ще е адекватен модел на света. Когато сме в някое от състоянията на графа, очакваме нещо да се случва. Очакваме някакви събития никога да не се случват в това състояние или обратното, задължително да се случват. Може някакво събитие да се случва с вероятност много по-голяма или много по-малка от средната вероятност за това събитие. Ако в никое от състоянията нищо не се случва, то модела ще е напълно неадекватен.

Това, което се случва когато се движим по стрелките на модела, ще наречем отпечатък на модела. При съвършеният модел имаме много ясен отпечатък. Там всяко състояние има точно определено наблюдение и точно определено множество от некоректните ходове. Ако вземем съкратения модел на съвършения, тогава някои стрелки ще липсват. Липсващата стрелка ще означава, че в определено състояние някакво събитие никога не се е случило (някакъв ход не е игран). Това също ще е отпечатък. Въпросът е как да го тълкуваме. Дали да предположим, че в това състояние това никога няма да се случи или че може да се случи, но с много малка вероятност.

Едно събитие може да зависи не само от това какво непосредствено сте видял и направил. То може да зависи и от по-далечната история. Например, какво сте видял на предишната стъпка.

Как ще търсим отпечатък? Ще го търсим чрез статистиката. Ще вземем съкратения модел (тоест, ще махнем всички стрелки и състояния, които до момента не са използвани). Ще преброим за всяка стрелка колко пъти е използвана. За всяко състояние и за всяко събитие ще преброим колко пъти то се е случило в това състояние. Разбира се, не можем да преброим за всяко събитие, защото събитията са безкрайно много. Ще броим само за някои по-важни събития. На базата на тази статистика ние можем да намерим отклонения от средното очаквано, които ще бъдат търсения отпечатък.

Нека да отбележим, че тази статистика може да я направи света, но за агента това би било много по-трудно. Света знае точно в кое състояние се намира и може да брои стрелките, докато агента може само да гадае. Дори и при съвършения модел агента може да не знае в кое състояние е бил (може да не знае кое е последното състояние, а дори и да знае, това не определя еднозначно предишните състояния).

Понякога агента не знае в кое състояние се намира в момента и това го научава в последствие. Това няма да е проблем, защото статистиката може да се събере със закъснение. Проблем е, когато агента никога не разбира точно в кое състояние е бил. Тогава проблема за събирането на статистиката е доста сложен.

Пълен модел

Да допуснем, че имаме един модел, в който оракулите α, β и χ не зависят от миналото. Този модел ще наречем пълен. Тоест, това е модел, който не може да се подобри. Всичко за миналото, което трябва да се запомни, вече сме го запомнили в текущото състояние.

Модела би могъл да се подобри, ако едно състояние го разделим на две и ако при определена история попадаме в едното или в другото с различна вероятност. Тоест, двете състояния да са различими спрямо миналото. При пълен модел двете състояния няма да са различими спрямо бъдещето, защото оракулите не завият от миналото. Тоест, това разделяне на състоянието на две става безпредметно.

Може да имаме пълен модел, който да бъде само с едно единствено състояние. В този случай света е ужасен и за бъдещето няма никакво значение какво се е случвало преди. В такъв свят няма никакво значение кое действие ще изберем. Разбира се, такъв свят е прекалено елементарен. Предполагаме, че световете, които ни интересуват, са много по-сложни и при тях намирането на съвършен или на пълен модел е на практика невъзможно.

Ние няма да се опитваме да разберем света напълно (да намерим съвършен модел) или да го разберем до нивото на някаква неразрешима случайност (да намерим пълен модел). Ние ще търсим много различни модели, които описват различни черти на света и обекти в света. Ще направим декартовото произведение на всичките тези модели и ще получим модел, който вероятно пак няма да е пълен, но ще описва света доста добре.

Това, че света е пълен се нарича „свойството на Марков“. В RL обикновено се предполага, че света има пълен модел. Тук ние предположихме нещо повече. Предположихме, че света има съвършен модел.

Ние няма да търсим пълни модели, а адекватни модели. Тоест, ще търсим модели, при които има някакъв отпечатък. Търсенето на пълен модел е една излишно амбициозна задача, която при по-сложните светове е абсолютно нерешима.

Видими събития

Събитие ще бъде някаква булева функция, която във всеки момент от времето е истина или лъжа. Първо ще кажем какво е видимо събитие. Това ще бъде събитие, което се вижда от историята (даже не от цялата история, а от нейния край).

Дефиниция: Видимо събитие ще бъде множество от локални истории. Това събитие ще бъде истина когато историята завършва с някоя локална история от множеството.

Пример за видимо събитие, това е какъв е последния ни ход. Пример за събитие, което не е видимо, това е дали един ход е коректен. Това се вижда от пълната история, но не се вижда от обикновената история. Това събитие ще го наречем полувидимо защото съдържа в себе си (като подмножество) едно видимо събитие. Видимото събитие, което се съдържа в него е следното: „този ход е некоректен и вече сме го пробвали“.

Друг пример за полувидимо събитие е изгрева. Него можете да го видите, но може и да го проспите. Това, че сте го проспали, не означава, че е нямало изгрев. Имало е, просто вие не сте го видели.

Пример за невидимо събитие е: настинах. Вие не виждате кога това събитие се случва, но в последствие може да установите, че то се е случило по различни признаци (като висока температура и други).

Досега множеството от събития, което използвахме за да намерим модел на света беше множеството от действията ни. Сега ще обобщим и ще използваме произволно множество от събития.

Event-Driven модел

Сега ще заменим действията със събития. Ще вземем едно множество от събития. Предполагаме, че това множество е малко, защото ако събитията са прекалено много модела ще стане прекалено сложен.

В новия модел стрелките вече няма да са по действия, а ще са по събития. При действията не беше възможно две действия да бъдат извършени едновременно. При събитията е напълно възможно две събития да се случат едновременно. Затова оракула α няма да се определя от действие, а от събития, при това, не от едно събитие, а от *events* – множество от събития. Когато *events* е празното множество, тогава никъде не отиваме и α ни връща същото състояние s_t. Когато в *events* има само едно събитие, тогава оракула избира една от възможните стрелки, които са по това събитие. Когато в *events* има две събития, тогава оракула трябва да реши дали да избере едно от двете (т.е. все едно, че едното събитие е затъмнило другото) или да предположи, че двете събития са се случили едно след друго. Тоест, да мине първо по стрелка отбелязана с първото и после да мине по още една стрелка, отбелязана с второто. Освен това оракула трябва да реши кое събитие да е първото и кое второто.

S – множество от вътрешните състояния на света.
s_t – текущото състояние на света.
E – множество от събития.
$G = <S, R>$ – тотален ориентиран граф (недетерминиран).
$R \subseteq S \times E \times S$
$\alpha(Past, s_t, events, Future) \rightarrow s_{t+1}$
$\beta(Past, s_{t+1}) \rightarrow v_{t+1}$
$\chi(Past, s_{t+1}, e) \rightarrow \{true, false\}$

Тук оракула χ определя не само кой ход е некоректен, а определя всички невидими събития от E. За полувидимите събития оракула определя тяхната невидима част. Видимите събития зависят само от *Past* и за тях е ясно как би работил оракула. Затова е излишно оракула да разпознава видими събития. Не искаме оракула да разпознава видими събития, защото пълен модел имаме когато оракулите не зависят от *Past*. За да не променяме дефиницията на пълен модел ще предполагаме че χ разпознава само невидимите събития, а видимите са ясни.

При event-driven модела пак можем да направим съкратения модел от стрелките, които са използвани и да преброим колко пъти са използвани. Пак света е този, който може точно да преброи, а агента може да направи само приблизителна преценка на тези бройки.

На базата на тази статистика може да намерим отпечатъка на модела и да преценим дали този модел е адекватен.

Модел с променливи

Естествено е към модела да добавим и променливи. В едно състояние едно събитие може да се случва известно време и после да престане да се случва. Това дали събитието в момента се случва е удобно да бъде представено като променлива. Тази променлива ще е локана, тоест, ще е свързана с определено състояние. Нищо не пречи да имаме и глобални променливи, които да са свързани с няколко състояния.

Пример за модел с променливи беше даден в [5]. Там имаше много врати (състояния) и всяка врата беше отключена или заключена. Тоест, на всяка врата съответстваше една променлива.

S – множество от вътрешните състояния на света.
Var – множество от променливи.
s_t – текущото състояние на света.
$eval_t$ – текущата оценка на променливите.
E – множество от събития.
$G = <S, R>$ – тотален ориентиран граф (недетерминиран).
$R \subseteq S \times E \times S$
$\alpha(Past, s_t, eval_t, events, Future) \rightarrow <s_{t+1}, eval_{t+1}>$
$\beta(Past, s_{t+1}, eval_{t+1}) \rightarrow v_{t+1}$
$\chi(Past, s_{t+1}, eval_{t+1}, e) \rightarrow \{true, false\}$

Тук от теоретична гледна точка нищо не правим, защото добавяйки променливи ние само увеличаваме броя на състоянията. Това се вижда и по-горе. Навсякъде където имахме състояние, сега има състояние и оценка на променливите.

Макар, че на теория добавянето на променливи нищо не променя, на практика се получава нещо много различно, защото ако имаме десет булеви променливи, то състоянията се увеличават 1024 пъти, което е много съществено увеличение. Ако се опитаме да нарисуваме ориентирания граф с толкова много състояния ще се получи нещо много сложно. В крайна сметка променливите описват отпечатъка (какво се случва в състоянието). За повечето променливи ние няма да имаме никаква идея каква е тяхната стойност.

Декартов модел

Както казахме, ще търсим различни модели, които да описват различни особености на света. Нашият модел на света, модела който сме открили, ще се състои от всичките тези модели. Може да си мислим, че той е декартовото произведение на всички тези модели.

Кое ще е състоянието, в което се намираме? Това ще са текущите състояния на всичките тези модели и текущите оценки на променливите в тези модели. Може да си мислим, че това е наредената енторка (tuple) от всичките тези текущи състояния и текущи оценки на променливи.

В момента вие сте в града, в който живеете. Денят е понеделник. Часът е 10 сутринта. Вие сте сит, защото закусихте добре. Всяко едно от тези четири изречения описва един event-driven модел и текущото състояние на този модел.

В момента вие виждате монитора. Това е обект (event-driven модел) и вие го виждате, тоест не сте в състоянието *outside*.

В момента вратата на вашата стая е отключена, а външната врата е заключена. Тоест, вие имате в главата си модел на сградата и имате идея за две от вратите дали са заключени. Тоест, знаете стойността на две от променливите на този модел.

Ще предполагаме, че никой от моделите, които сме открили не е пълен, защото ако има един пълен, то другите модели ще са излишни. Възможно е декартовото произведение на много непълни модели да даде пълен модел, но ако света е достатъчно сложен, това ще е малко вероятно.

Агенти и обекти

Трябва да правим разлика между агенти и обекти. Това са две различни неща.

В предишни статии [7] говорихме за агенти. Когато някой нещо променя, то този някой е агент и ние трябва да предвидим следващите му действия, да решим дали ни е враг или съюзник, да се опитаме да се разберем с него.

Какво е човека – агент или обект. От една страна той е агент, защото променя света. Може нещо да премести, може нещо да открадне. От друга страна, той е обект, защото ние го разпознаваме. Като го видим, като го чуем по телефона, когато някой спомене името му.

Заключение

В тази статия разгледахме връзката между събитията и тяхното следствие. (Следствието наричаме отпечатъка на event-driven модела.) Въпросът е, кога се случва събитието? Например, кога с случва изгрева? Когато слънцето се показва малко, когато се показва наполовина или когато напълно се показва? Тук избрахме един момент, в който приемаме, че събитието се е случило и това е момента, в който сме разбрали, че то се е случило. Въобще не е задължително, следствията от това събитие да започнат точно в този момент. Например, когато слънцето изгрее става светло веднага, дори става светло още преди да е изгряло. От друга страна, става топло, но не веднага, а много по-късно. Тоест, когато търсим отпечатъка, може да има следствие от събитието или следствие от това, че сме попаднали в определено състояние на модела, но не трябва да очакваме това следствие да се появи веднага. Трябва да допускаме следствието да е малко изместено във времето. Освен това следствието (отпечатъка) няма да е свързано само със състоянията на модела. Може да имаме следствие и от самото събитие. Например при изгрева небето става червено и това е около момента на изгрева. Това не е свързано със състоянието преди или със състоянието след изгрева (тоест, с нощта или с деня).

Тоест, когато събираме статистика трябва да отчитаме колко близо сме до събитието и да търсим отпечатък (особености), както в периодите между събитията, така и около самите събития. Отпечатъка е особеност, която се проявява в определено състояние на модела, но може да се проявява и когато минаваме по някоя от стрелките на модела.

Отпечатъка няма да се състои само от събития. Освен събитията, ще имаме и тестове, ще имаме и обекти. Тестовете са специален тип събития, за които ще кажем по-надолу. Появата на обект също е събитие и това събитие може да ни помогне да дефинираме отпечатъка на един event-driven модел. Тоест, за дефинирането на обект ние може да използваме други обекти. Например, в една стая виждаме котка и това е стаята с котката. Ето как обекта котка ни помага да определим и да запомним една стая. Разбира се, котката е обект, който има свойството да се мести. Може да се наложи да я търсим из цялата къща. Ако котката има свойството единственост, тогава ако сме я намерили в една стая няма да я търсим повече по другите стаи. Може котката да не стои в една стая, но в различните стаи да има различна вероятност да видим котката. Тази различна вероятност също може да бъде отпечатък на модела (в случая модела е къщата).

Видяхме, че каквито и събития да си вземем и какъвто и ориентиран граф да изберем, то това е модел на света. Лошото е, че ако този модел е избран произволно, то той почти сигурно ще се окаже неадекватен. Тоест няма да има никакъв отпечатък или нищо особено (нищо интересно) няма да се случва в състоянията му.

В някои случаи (като фигура 2) модела ще има отпечатък, но този отпечатък ще се дължи само на оракула α и ние няма как да използваме този отпечатък (в случая да предскажем следващото събитие), защото няма да има как да познаем в кое състояние сме.

Това показва, че задачата за откриването на адекватен модел на света е много сложна. За целта трябва да търсим особености. Трябва да наблюдаваме различни събития и да забележим когато някоя особена комбинация от събития се случи. Например, черна шапка и щръкнала коса – това трябва да е колегата Джон.

Малко са събитията, които можем да наблюдаваме постоянно. Повечето събития ние ги наблюдаваме само понякога. Такива събития наричаме тестове [5]. Тестовете са особено важни за откриването на адекватни event-driven модели.

Следващата статия ще бъде посветена на тестовете и ще видим как с тяхна помощ можем ефективно да откриваме event-driven модели.

References

[1] Richard Sutton, Andrew Barto (1998). Reinforcement Learning: An Introduction. *MIT Press, Cambridge, MA (1998).*

[2] Richard Sutton (2008). Fourteen Declarative Principles of Experience-Oriented Intelligence. www.incompleteideas.net/RLAIcourse2009/principles2.pdf

[3] Johan Åström. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications. 10: 174–205.*

[4] Apostolos Burnetas, Michael Katehakis. (1997). Optimal Adaptive Policies for Markov Decision Processes. *Mathematics of Operations Research. 22 (1): 222.*

[5] Dimiter Dobrev (2017). How does the AI understand what's going on. *International Journal "Information Theories and Applications", Vol. 24, Number 4, 2017, pp.345-369.*

[6] Dimiter Dobrev (2018). Minimal and Maximal models in Reinforcement Learning. *August 2018. viXra:1808.0589.*

[7] Dimiter Dobrev (2008). The Definition of AI in Terms of Multi Agent Systems. *December 2008. arXiv:1210.0887.*