Универсална интелигентност: Дефиниция на агента на Маркус Хутер

Universal Intelligence:
A Definition of Machine Intelligence

Shane Legg

IDSIA, Galleria 2, Manno-Lugano CH-6928, Switzerland

shane@vetta.org www.vetta.org/shane

Marcus Hutter

RSISE @ ANU and SML @ NICTA, Canberra, ACT, 0200, Australia

marcus@hutter1.net www.hutter1.net

December 2007

Съдържание

1. Човешка интелигентност
2. Дефиниция на машинна интелигентност
3. Дефиниция на тест на интелигентност за машина
4. Анализ и обобщение

Лекцията е по статията на M.Hutter и S.Legg, с въведение, забележки и примери от мен.

Кой е по-умен?

- Кучето Рекс, котката Нора или папагалът Коко?
- Отличникът на класа („зубъра“) или „душата на компанията“?
- Христо Стоичков или акад. Благовест Сендов?
- Един и същ човек на 1, 3, 12, 25, 50, 70 години?
- Шах-компютърът Deep Blue или мишката Мики?
- Мъжете или жените?
- Роботът Азимо или кучето робот Аибо?

Типове интелигентност

- „Течна“ (флуидна) и „Кристализирана“
- Пространствена
- Езикова
- Емоционална (EQ)
- Математическа
- Музикална
- Мускулна/Спортна
- ... Не може да мерим всичко с един аршин?... Но...

Обща абстрактна дефиниция на интелигентност е...

Необходима за построяването на зародиш на универсален самоусъвършенстващ се разум.

Предполага се, че неокортекстните колони (стълбчета) в мозъка на бозайниците и човека са „градивни елементи“ на техните висши когнитивни способности.

Тестове за интелигентност

- Francis Galton – времена за реакция,
- Тест на Бине, 1905 г. - за деца, 30 въпроса от няколко типа, нарастваща сложност във всеки тип: *именуване на части от тялото, сравняване, броене на монети, запомняне на цифри и определения.* Добре предсказвал бъдещия успех в училище.
- Тест тип Станфорд-Бине – приспособен за САЩ, Бине + американски тестове за новобранци Алфа А, Алфа Б
- Дейвид Уелшър 1950-те – добавя невербални въпроси; тестове за различни възрастови групи
- Прогресивни матрици – чисто визуални тестове. (Тестовете на Менса)

IQ

- Статистическа характеристика – колко % от хората в дадена група, възрастова група или в света имат по-висок/по-нисък резултат на даден тест.
- Учудващо стабилен
- Гаусово разпределение, 100 – среден за дадена „ментална възраст“ или популация. Десет-годишно дете със способности на 12-годишно има IQ = 120.
- Менталната възраст е оспорвано понятие

Образователни тестове

- Изрични или неизрични
- Образование в детската градина, игри
- Образователен минимум за дете преди да влезе в училище
- Сложността на предметите, с които учникът трябва да се справи

Интелигентност на животните

- Различно развити сетива спрямо човешките *(обонянието е водещо сетиво за много животни, при нас основното е зрението).*
- Различни тестове за различно интелигентни животни.
- По-прости – краткотрайна и дълготрайна памет, образуване на асоциации между стимули, схващане на прости зависимости и предсказване, броене и общуване.
- По-сложни – служат си с измама, имитират, разпознават се в огледало (Mirror-test).
- Как да ги накараме да направят теста? С награди... Насочване на поведението чрез учене с подсилване.

Желани свойства на IQ тест

- Да е повторяем.
- Да не се влияе от културни особености

  (Тест на чужд език със запълване на липсващи думи в редки книжовни думи на чуждия език.)
- Да дава действителни мерки спрямо това което твърди, че измерва.
- Да предсказва, напр. бъдещите академични резултати.
- Да е лесен за проверка.

Статични и динамични тестове

- Статични - стандартните добре познати тестове са статични, измерват знания и/или способност за решаване на фиксирани въпроси/задачи.
- Динамични - измерват способността за обучение и приспособяване.

Теории за човешката интелигентност

- Една обща или много различни способности?
- „Множество фактори“ - 7 първични умствени способности: езиково разбиране, говорене, работа с числа, пространствено изобразяване, асоциативна памет, скорост на възприемане, разсъждение.
- „Три арки“ - аналитична, творческа и практическа
- „Многостранна интелигентност“ - езикова, музикална, логико-математическа, пространствена, телесно-двигателна, вътреличностна и междуличностна
- **G-фактор – обща интелигентност, различните компоненти са статистически взаимосвързани**
- „Течна“ (флуидна) и „Кристализирана“

Определения за човешкия ум

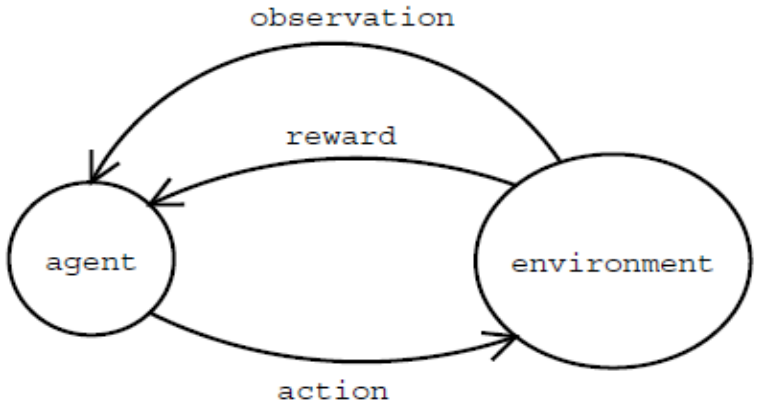
- Hutter, p.11 - 14

Определение за машинен разум

- Агент, среда, цели - Agent, Environment, Goals.
- Комуникация между средата и агента:
  - агентът има възприятия от средата
  - агентът въздейства върху средата
- Откъде да знае каква е целта – **вградена** или задавана с команди (**чрез език**).
- **Награда – универсален метод, учене с подсилване.**
- Сигнал, който показва дали агентът се справя – целта на агента е да максимизира наградата.

Определение за машинен разум

- A:={ляво, дясно, горе, долу} – агентът праща информация към средата (взаимодействия), пространство на действията (Actions)
- P - средата връща сигнали в пространство на възприятията (Perceptions)
- R – пространство на наградата (Reward) [0..1]
- P:={(студено, 0.0), (топло, 1.0), (горещо 0.3))}
- O – наблюдение (observation)
- o1r1a1o2r2a2o3r3a3o4 ...
- Наблюдение, награда, действие...



Функция на поведението на агента

- П (пи) – функция на агента, за която историята е вход, и тя връща следващото действие.
- Детерминистична - връща винаги вероятност 1 за дадено действие при дадена история (държи се по един и същи начин).
- Вероятностна (недетерминирана): връща вероятност между 0..1:

  П(a3|o1r1a1o2r2) - вероятността да се извърши действие a3 в третия цикъл, ако историята до момента е: o1r1a1o2r2
- Поведението на човека може да се екстраполира и приеме като такава функция.

Функция на средата и мярка за успех

- μ (мю) – за всяко k, вероятността от OkRk зависи от миналото:

  μ(**Okrk|o1r1a1o2r2a2...ak-1rk-1ak-1**)

  Мярка за успех на агент

- A1 – бързо намира начин да получава награда от 0.9 и прави същото действие, което му я дава - **в началото е по-успешен.**
- A2 – известно време търси докато открие 1.0 **(в началото се лута и е по-неуспешен, но след това е по-успешен). Период за планиране на действията.**
- **„Изследване или използване“ -** Explore vs. Exploit

Формално определение за разум

При наличие на много хипотези, които са несъгласувани с данните, да се предпочете най-простата.

Това се приема за „рационално“
Бръсначът на Окам
Тестове за интелигентност...

Опасни „бъгове“ в средата и агента

- От гледна точка на агента, неточен модел на средата може да бъде оптимален, ако грешките нямат отношение към получаваната награда.
- Бръсначът на Окам се отнася до сложността на хипотезата/теорията, а не до трудността за изпълнение на добра стратегия.
- За да се отличат агентите, които правилно ползват бръснача на Окам, трябва да се измерва сложността на средата, а не трудността за постигане на целта.

Опасни „бъгове“ в средата и агента

- Средата трябва да е достатъчно сложна
- Ако наградата е винаги 1, независимо от това че връзката между действията и наблюденията е сложна, агентът няма нужда да търси оптимална стратегия.

При човека - удоволствието не трябва да се получава прекалено лесно и да бъде еднообразно – пристрастяване, наркотици и пр.

Сложност на Колмогоров

- K(x) := min{ *L*(p): U(p) = x}
- p – двоичен низ, програма
- L(p) – дължина на програмата
- U – универсална машина на Тюринг, *образцова машина*
- *K(x) – дължина на най-кратката програма p, която изчислява низа x чрез образцовата машина U.*
- *Няма къси програми за за дълги случайни низове*
- K е почти независим от избора на U

Реален пример

- 1 милиард нули – цикъл, който пише N/4 пъти 32-битов шаблон в паметта на 32-битов процесор x86.

```
INIT:  MOV edx, start_addres  //edx – начален адрес
       MOV  eax, edx           //eax ще помни...
       ADD   eax, N            //крайният адрес
       MOV  ecx, pattern       //шаблон
CYC:MOV  [edx], ecx            //шаблон --> памет
       ADD   edx, 4            //следваща клетка
       CMP  edx, eax           //проверка за край
       JNG   CYC      //ако не е свършил - продължава...
```

Сложност на средата

- $\mu_1, \mu_2, \mu_3, \ldots$ - различни среди [виртуални светове, вселени]
- $\langle i \rangle$ - низ, генериран от програма с тази сложност
- $K(\mu_i) := K(\langle i \rangle)$ - сложност на средата
- Приемаме, че всеки допълнителен бит в описанието, намалява вероятността на средата наполовина.
- $2^{-K(\mu)}$ ==> Алгоритмична вероятност на пространството от среди.

Индукция, вероятностното разпределение – дефиниране на универсални учещи агенти с доказеума оптимална производителност.

Сложност на средата

- $\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi$ - очаквана успеваемост на агент, или УИР на агент по М. Хутер (Υ - ипсилон)
- E – пространство от среди (Environments)
- V – „value function“ - получена награда от средата
- *Вътрешната работа на агента не е от значение – всяко „нещо“, което може да извежда и да получава информация за да постига цели.*
- K обаче не е изчислима, може да се изчисли само с приближение

Υ

Видове агенти според средата

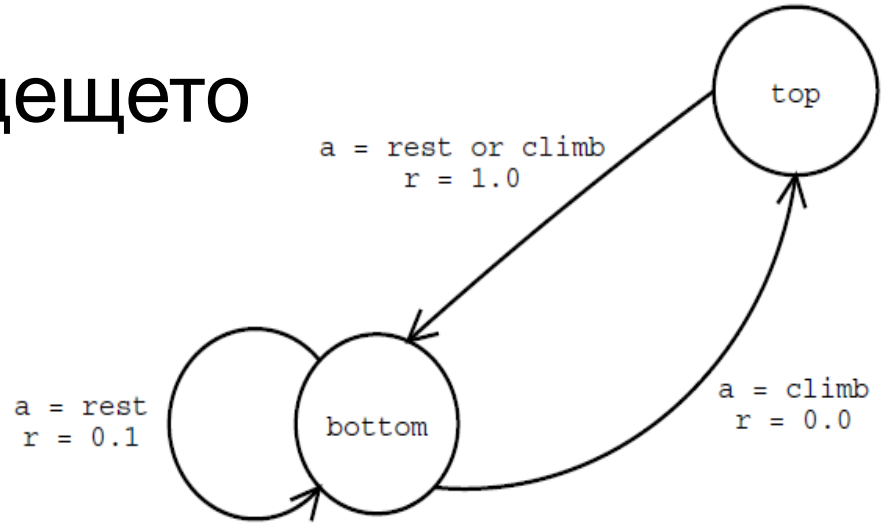
- Случаен $\Upsilon(\pi^{\mathtt{rand}})$
- Много специализиран – Deep Blue.
- Универсален, но прост.
- Прост агент с по-дълга история.

$$\mathcal{R} = [0,1] \cap \mathbb{Q},\ \mathcal{A} = \{\mathtt{up}, \mathtt{down}\} \text{ and } \mathcal{O} = \{\varepsilon\}$$

$$\mu^{\mathtt{alt}}(o_k r_k | o_1 \ldots a_{k-1}) := \begin{cases} 1 & \text{if } a_{k-1} \neq a_{k-2} \wedge r_k = 2^{-k}, \\ 1 & \text{if } a_{k-1} = a_{k-2} \wedge r_k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Видове агенти според средата

- Прост предсказващ агент: изчислява наградата в бъдещето



- Много умен агент
- Свръхумен агент AIXI/съвършен агент

$$\bar{\Upsilon} := \max_{\pi} \Upsilon(\pi) = \Upsilon\left(\pi^{AIXI}\right).$$

Нормализирана награда

$$V_\mu^\pi := \mathbf{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1.$$

Сборът от наградите за целия измерван период трябва да е нормализиран (ограничен от 1), за да се избегне отклонение заради дължината на периода.

Свойства на универсалния разум

- Случайност на редица по Мартин-Льоф – когато няма „значителен ред“, т.е. Редицата не може да бъде компресирана. Подобно на сложността K.
- Zip, RAR, 7zip... ОК, но може да има някаква много сложна хитра зависимост, за която не подозираме, и може да даде голяма компресия.

  Случайна редица? **1, 8, 2, 4, 5, 2, 8, 9, 6**
  Но... 52896/1824 = 29 (просто число)

  „Вселена и Разум 4“, Т.Арнаудов 2004

- Една поредица е „случайна“ спрямо способностите за откриване на зависимости на оценителя на поредицата.

Свойства на мярката за УИР

Υ
ипсилон

- Валидна, смислена, информативна
- Широкообхватна $\pi^{\texttt{rand}}$, $\pi^{\texttt{basic}}$, $\pi^{\texttt{2back}}$ and $\pi^{\texttt{2forward}}$
- Обща
- Безпристрастна
- Фундаментална (изчисление, информация, сложност)
- Формална $\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi$
- Обективна
- Универсална – не антропоцентрична
- Практична – тест с приближени стойности

Неформални определения на разум

- „Умствени способности за поддържане на успешен живот...“ - K. Warwick, цитиран в [Aso03], 2003.
- Справяне с широк обхват от задачи е емпирично определение на „интелигентност“ - H. Masum 2002.
- Изчислителната част на способността за постигане на цели в света. В различни степени интелигентността е присъща на хората, много животни и някои машини. - J. McCarthy 2004.
- Всяка система, която поражда адаптивно поведение в разнообразни среди може да бъде наречена разумна. D. Fogel, 1995.
- **Способността на система да се държи по подходящ начин в неопределена среда, където подходящите действия увеличават вероятността за успех, а успехът е постигането на поведенчески подцели, които подкрепят главната цел. (...в голямо разнообразие от обстоятелства) - J.S. Albus 1991 – подобно на дефиницията на М. Хутер**

Неформални определения на разум

- „От разумните системи се очаква да работят добре в много различни среди, като техният ум позволява да увеличават вероятността за успех дори ако не е възможно да се постигне пълно знание за средата. Функционирането на разумната система може да се разглежда отделно от средата и конкретната ситуация, включително целта.“ - R. R. Gudwin '00. *Gudwin изисква определен начин по който системата работи, не просто черна кутия, докато за Хутер вътрешните принципи на действие са без значение.*
- „Определяме две гледни точки върху разума на една машина: (1) вроден, изразена в определена наследена сложност на информационното съдържание на системата, и (2) действителен – той се изразява в успешността (постигащо цели) поведение в сложна среда.“ - J. A. Horst'02
- „...способността да решава сложни задачи“. M. Minsky [Min85]
- „Постигане на сложни цели в сложни среди.“ B. Goertzel [Goe06]

Неформални определения на разум

“. . .във всяка реална ситуация се държи подходящо според възможностите на машината и се приспособява според изискванията на средата, с някои ограничения в скоростта и сложността.” A. Newell и H. A. Simon [NS76]

„[Разумният агент прави това, което] е подходящо за обстоятелствата и целите му, гъвкав е в променящи се среди и променящи се цели, учи от опита и избира правилно, спрямо ограниченията на възприятията и крайната му изчислителна мощ.“ - D. Poole [PMG98]

„Способността да използваш оптимално ограничени средства – включително времето – за постигане на цели.“ R. Kurzweil [Kur00]
„Способността на обрабоващ информация агент да се приспособява към средата при недостатъчно знание и налични средства.“
Ограничеността на ресурсите е важна от практическа гледна точка.

Тестове за изкуствен разум

- **Тест на Тюринг –** най-стар, чат между човек и машина

  Много недостатъци – наивен, машината трябва да се прави на човек или да заблуди човека, че не е машина (може да е много по-умна и бърза от човек).

  Loebner Prize – за чат-ботове

- **Тест за компресиране** -

  Hutter Prize – 100 MB корпус от Уикипедия.

- **Тест на езиковата сложност –** брой изполвани думи, дължина на изречения, видове отговори, синтактична сложност и др.

Тестове за изкуствен разум

- **Множество познавателни способности –** IBM *Joshua Blue*, и Adaptive AI *a2i2*
  Различни тестове - езикови, за общуване, асоциативни, обучение с различна трудност. „Бебешки Тюрингов тест“.
- **„Образователен тест“** (Т. Арнаудов) – спрямо дефинираните норми за когнитивно развитие на човека на различна възраст/образователна степен.
- **Състезателни игри –** напр. шах коефициент.
- **Психометрични тестове** – прилагане на човешки тестове за интелигентност, Bringsjord и Schimanski.
  Критика: Машината може да е специализирана точно за този тип тестове, и да не е универсална.

Тестове за изкуствен разум

- **C-Test (Complexity Test) –** предсказване на следващ символ. Използва формалната мярка за сложност на Левин Kt – подобна на K, но изчислима, за разлика от K - машината наТюринг трябва да може да се симулира за линейно време.
  Недостатък - статичен тест.

Sequence Prediction Test

Complexity	Sequence	Answer
9	a, d, g, j, _ , . . .	m
12	a, a, z, c, y, e, x, _ , . . .	g
14	c, a, b, d, b, c, c, e, c, d, _ , . . .	d

Sequence Abduction Test

Complexity	Sequence	Answer
8	a, _ , a, z, a, y, a, . . .	a
10	a, x, _ , v, w, t, u, . . .	y
13	a, y, w, _ , w, u, w, u, s, . . .	y

Сравнение на тестове за изкуствен разум

- **Валиден –** трябва да измерва именно разумността
- **Информативен –** резултатът трябва да е скаларен, може би вектор или абсолютна стойност, за да може да се сравнява.
- **Широкообхватен –** от много ниска до много висока
- **Всеобщ –** да е приложим както за ума на най-прости същества, така и за свръхчовешки разум
- **Динамичен –** да отчита способността за учене и приспособяване.
- **Безпристрастен –** да не дава предимство на определена култура или вид същество

Сравнение на тестове за изкуствен разум

- **Фундаментален –** да не се променя с напредък на технологиите и познанието
- **Формален –** максимално прецизно определен, ако е възможно описан на формален математически език.
- **Обективен –** да не зависи от субективни (човешки) оценки.
- **Напълно определен –** всички аспекти и мерки да са напълно определени.
- **Универсален –** да не бъде антропоцентричен.
- **Практичен –** да се измерва автоматично и бързо.

Сравнение на тестове за машинен разум

Intelligence Test	Valid	Informative	Wide Range	General	Dynamic	Unbiased	Fundamental	Formal	Objective	Fully Defined	Universal	Practical	Test vs. Def.
Turing Test	•	·	·	·	•	·	·	·	·	•	·	•	T
Total Turing Test	•	·	·	·	•	·	·	·	·	•	·	·	T
Inverted Turing Test	•	•	·	·	•	·	·	·	·	•	·	•	T
Toddler Turing Test	•	·	·	·	•	·	·	·	·	·	·	•	T
Linguistic Complexity	•	●	•	·	·	·	·	•	•	·	•	•	T
Text Compression Test	•	●	●	•	·	•	•	●	●	●	•	●	T
Turing Ratio	•	●	●	●	?	?	?	?	?	·	?	?	T/D
Psychometric AI	●	●	•	●	?	•	·	•	•	•	·	•	T/D
Smith's Test	•	●	●	•	·	?	●	●	●	·	?	•	T/D
C-Test	•	●	●	•	·	●	●	●	●	●	●	●	T/D
Universal Intelligence	●	●	●	●	●	●	●	●	●	●	●	·	D

Table 1: In the table ● means "yes", • means "debatable", · means "no", and ? means unknown. When something is rated as unknown that is usually because the test in question is not sufficiently specified.

Test vs. Def. Finally, we note whether the proposal is more of a test, more of a definition, or something in between.

Връзки

-
- Hut07b] M. Hutter. Universal algorithmic intelligence: A mathematical top→down approach. In *Artificial General Intelligence*, pages 227–290. Springer, Berlin, 2007
- [Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. 300 pages, http://www.hutter1.net/ai/uaibook.htm..
-
-
-

Връзки

The genesis of this work lies in Hutter's universal optimal learning agent, AIXI, described in 2, 12, 60 and 300 pages in [Hut01b, Hut01a, Hut05, Hut07b], respectively. In this work, an order relation for intelligent agents is presented, with respect to which the provably optimal AIXI agent is maximal. The universal intelligence measure presented here is a derivative of this order relation. A short description of the universal intelligence measure appeared in [LH05], from which two articles followed in the popular scientific press [GR05, Fi´e05]. An 8 page paper on universal intelligence appeared in [LH06b], followed by an updated poster presentation [LH06a]. In the current paper we explore universal intelligence in much greater detail, in particular the way in which it relates to mainstream views on human intelligence and other proposed definitions of machine intelligence.