Подход к генерации оптимального действия при взаимодействии системы с внешней средой*

Олег Волков

Редакция 0.82 от 8 мая 2012 г.

Аннотация

В статье приводится описание алгоритма работы системы, пошагово взаимодействующей с внешней средой посредством приёма и выдачи сигналов. Целью работы системы является получение максимального количества стимула (энергии или некоторого ресурса) из внешней среды. В ключевой точке алгоритма используется генератор случайных чисел. Понятиями, на которых основан алгоритм, является цикл в графе переходов между состояниями системы и средняя величина прироста стимула за шаг такого цикла. Также рассматривается вариант построения системы, составленной из двух подсистем, первая из которых определяет текущее состояние внешней среды, а вторая на основе этого состояния генерирует оптимальное действие. Идея алгоритма может быть использована в системах управления или интерактивного взаимодействия с человеком, функционирующих в режиме реального времени.

1 Описание алгоритма

Каждый шаг $t \in \overline{1, n_{ev}}$ взаимодействия системы с внешней средой описывается тройкой $\left\langle s_{in}^{(t)}, \Delta E^{(t)}, s_{out}^{(t)} \right\rangle$ и состоит в приёме *информационного сигнала* $s_{in}^{(t)} \in \overline{1, n_{in}}$, являющегося приближением к состоянию внешней среды на шаге t, величины прироста стимула (энергии или некоторого ресурса) $\Delta E^{(t)}$ и в выдаче во внешнюю среду *сигнала действия* $s_{out}^{(t)} \in \overline{1, n_{out}}$. Целью работы системы является максимизация значения $E = \sum_{t=1}^{n_{ev}} \Delta E^{(t)}$.

Состоянием системы является n-грамма[1] длины N или меньшей, если в виду малых значений t количество произошедших событий есть недостаточным для формирования n-граммы длины N. Пусть $Sx_i \in \overline{1, n_{in}}$, $Ay_i \in \overline{1, n_{out}}$, тогда n-граммы вида $\langle ...Sx_{N-3}, Ay_{N-2}, Sx_{N-1}, Ay_N \rangle$ соответствуют состояниям выполнения действия, а n-граммы вида $\langle ...Ay_{N-3}, Sx_{N-2}, Ay_{N-1}, Sx_N \rangle$ — состояниям выбора действия. Последовательность перехода состояний

$$\left\langle ...Sx_{N-1}, Ay_N \right\rangle \to \left\langle ...Sx_{N-1}, Ay_N, s_{in}^{(t)} \right\rangle \to \left\langle ...Sx_{N-1}, Ay_N, s_{in}^{(t)}, s_{out}^{(t)} \right\rangle$$

соответствует одному шагу t, причём состоянию $\left\langle ...Sx_{N-1}, Ay_N, s_{in}^{(t)} \right\rangle$ соответствует подшаг $t + \frac{1}{2}$.

Цикл в графе переходов между n-граммами, отвечающими состояниям выбора действия, — ключевое понятие, используемое в алгоритме. Тип циклов определяется парой $\langle h, z \rangle$, где n-грамма h выбора действия задаёт начальное и конечное состояние циклов, $z \in \overline{1, n_{out}}$ задаёт направление циклов из начального состояния h.

В области памяти, соответствующей каждой n-грамме h выбора действия, хранится вектор

$$\left\langle \alpha^{(h)}, t_0^{(h)}, E_0^{(h)}, \psi^{(h)} \right\rangle$$

статуса для состояния выбора действия. В этом векторе $\alpha^{(h)} \in \overline{0, n_{out}}$ — сигнал действия, выданный непосредственно перед тем, как был осуществлён последний выход из узла h графа переходов (или 0, если такого сигнала ещё не было), который используется для определения направления цикла в узле h при последующем возврате в этот узел, $t_0^{(h)}$ — шаг последнего выхода из узла h, $E_0^{(h)}$ — стимул на шаге $t_0^{(h)}$, $\psi^{(h)}$ представляет собой таблицу соответствий возможных сигналов действия z векторам

$$\left\langle \nu^{(h,z)}, \omega^{(h,z)}, H^{(h,z)} \right\rangle$$

[1] N-грамма — это подпоследовательность длины N некоторой последовательности. Последовательностью в данном случае выступает история принятых и выданных сигналов $s_{in}^{(1)}, s_{out}^{(1)}, s_{in}^{(2)}, s_{out}^{(2)}, s_{in}^{(3)}, s_{out}^{(3)}, ...$ При $N = 2$ n-граммы представляют собой пары из двух соседних элементов.

статистики циклов, начальными значениями которых являются нулевые векторы. В векторах статистики циклов $\nu^{(h,z)}$ — количество циклов типа $\langle h,z\rangle$, встретившихся с момента начала работы системы, $\omega^{(h,z)}$ — суммарная длина встретившихся циклов типа $\langle h,z\rangle$, $H^{(h,z)}$ — суммарный прирост стимула на протяжении циклов типа $\langle h,z\rangle$.

При получении информационного сигнала $s_{in}^{(t)}$ осуществляется переход в состояние выбора действия, соответствующее n-грамме h длины $\min\{2t-1,N\}$. Если $2t-1<N$, то обработка информационного сигнала завершается. Иначе, если $\alpha=\alpha^{(h)}\neq 0$, выполняется обновление компонент вектора статистики циклов типа $\langle h,\alpha\rangle$, включающее инкремент количества $\nu^{(h,\alpha)}$ таких циклов, увеличение суммарной длины $\omega^{(h,\alpha)}$ циклов на разницу между текущим шагом t и шагом $t_0^{(h)}$ начала цикла и увеличение суммарного прироста стимула $H^{(h,\alpha)}$ на протяжении циклов на разницу между текущим стимулом $E^{(t)}=\sum_{i=1}^{t}\Delta E^{(i)}$ и стимулом $E_0^{(h)}$ на шаге $t_0^{(h)}$:

$$\begin{aligned}\nu^{(h,\alpha)} &:= \nu^{(h,\alpha)}+1;\\ \omega^{(h,\alpha)} &:= \omega^{(h,\alpha)}+t-t_0^{(h)};\\ H^{(h,\alpha)} &:= H^{(h,\alpha)}+E^{(t)}-E_0^{(h)}.\end{aligned}$$

Если $2t-1<N$, то в качестве сигнала действия $s_{out}^{(t)}$ выбирается случайный сигнал. В обычном случае, когда $2t-1\geq N$, сигнал действия $z\in\overline{1,n_{out}}$ в состоянии h выбирается с вероятностью, пропорциональной

$$F(h,z,t)=\exp\left(\frac{H^{(h,z)}}{\omega^{(h,z)}}K(h,z,t)\right),$$

где $H^{(h,z)}/\,\omega^{(h,z)}$ — средний прирост стимула за шаг циклов типа $\langle h,z\rangle$, $K(h,z,t)$ — множитель, при большем значении которого увеличивается вероятность выбора сигнала действия z, соответствующего большему значению $H^{(h,z)}/\,\omega^{(h,z)}$. Использование экспоненциальной формы для функции относительной вероятности, которой соответствует распределение Больцмана вероятности перехода в одно из возможных состояний системы, показало свою эффективность при решении задач глобальной минимизации как, например, в алгоритме, приведенном в [2].

После выдачи сигнала действия $s_{out}^{(t)}$ системой, если $2t-1\geq N$, выполняются присваивания

$$\begin{aligned}\alpha^{(h)} &:= s_{out}^{(t)};\\ t_0^{(h)} &:= t;\\ E_0^{(h)} &:= E^{(t)}.\end{aligned}$$

По завершении обработки сигнала действия осуществляются переход в состояние выполнения действия и инкремент номера шага t.

Множитель $K(h,z,t)$ может быть определён различными способами, например, следуя подходу, применяемому в алгоритме модельной “закалки” (simulated annealing algorithm), когда множитель растёт с увеличением t. В данной статье приведен собственный способ определения этого множителя.

Максимизируемой величиной является стимул, полученный системой за время своей работы, или, иначе, средний прирост стимула за шаг работы системы, равный $E^{(t)}/\,t$, который приближается к среднему приросту стимула за шаг цикла с увеличением t. Величина $E^{(t)}/\,t$ будет максимизироваться, если частота повторения циклов типа $\langle h,z\rangle$ будет тем выше, чем отношение среднего прироста стимула за шаг этих циклов к величине $\left|E^{(t)}\right|/\,t$ будет больше. Величина $t/\left|E^{(t)}\right|$ является коэффициентом для перевода величины $H^{(h,z)}/\,\omega^{(h,z)}$ в область значений w_z, не зависящих от масштаба средних приростов стимула за шаг встретившихся циклов:

$$w_z=\frac{tH^{(h,z)}}{\left|E^{(t)}\right|\omega^{(h,z)}}.$$

В случае, когда $w_z=1$ при

$$\frac{H^{(h,z)}}{\omega^{(h,z)}}=\frac{\left|E^{(t)}\right|}{t}$$

для некоторого z, и $w_i=0$ при $H^{(h,i)}=0$ для $i\in\overline{1,n_{out}}\setminus\{z\}$, выбор действия z будем осуществлять с вероятностью p_c, а выбор каждого действия i будем осуществлять с вероятностью $(1-p_c)/\,(n_{out}-1)$. Если в прошлый раз в состоянии h действие z было выбрано с вероятностью p_c, то с вероятностью $p_f=p_c^2$ выбранное в этот раз действие будет являться действием z и совпадать с действием, выбранным прошлый раз в состоянии h.

Если система сделала цикл типа $\langle h,z\rangle$ длиной κ подшагов и оказалась в состоянии h, тогда точное повторение последовательности сигналов, соответствующих подшагам цикла, снова приведёт систему в состояние h. Пусть имеем некоторую последовательность сигналов, тогда если сгенерировать вторую последовательность сигналов такой же длины, выбирая в каждой позиции второй последовательности с вероятностью p_f сигнал, который

был в этой позиции в первой последовательности, то обе последовательности будут иметь общий префикс длины ровно κ с вероятностью $p_r(\kappa) = p_f^{\kappa}(1-p_f)$, $\sum_{i=0}^{\infty} p_r(i) = 1$. Если зафиксировать значение p_f, то средняя длина общего префикса будет равна

$$\bar{\kappa} = \sum_{i=0}^{\infty} i p_r(i) = \frac{p_f}{1-p_f},$$

откуда $p_f = \frac{\bar{\kappa}}{\bar{\kappa}+1}$,

$$p_c = \sqrt{\frac{\bar{\kappa}}{\bar{\kappa}+1}}.$$

Вероятности p_c соответствует вероятность попадания равномерно выбранной случайной точки в отрезок длины x, за которым следуют $n_{out}-1$ отрезков длины $\exp 0 = 1$. Решая уравнение

$$\frac{x}{x+n_{out}-1} = \sqrt{\frac{\bar{\kappa}}{\bar{\kappa}+1}},$$

получаем

$$x = \sqrt{\bar{\kappa}}\left(\sqrt{\bar{\kappa}} + \sqrt{\bar{\kappa}+1}\right)(n_{out}-1).$$

Обеспечить равенство $F(h,z,t) = x$ при $w_z = 1$ и добиться, чтобы при изменении w_z значение функции $F(h,z,t)$ изменялось по степенному закону, соответствующему выбранной экспоненциальной форме для этой функции, можно, определив функцию как $F(h,z,t) = B(h,z)^{w_z}$, где $B(h,z) = x$:

$$\begin{aligned} F(h,z,t) &= B(h,z)^{\frac{tH^{(h,z)}}{\left|E^{(t)}\right|\omega^{(h,z)}}} = \left(\sqrt{\bar{\kappa}}\left(\sqrt{\bar{\kappa}} + \sqrt{\bar{\kappa}+1}\right)(n_{out}-1)\right)^{\frac{tH^{(h,z)}}{\left|E^{(t)}\right|\omega^{(h,z)}}}; \\ K(h,z,t) &= \frac{t}{\left|E^{(t)}\right|}\ln B(h,z) = \frac{t}{\left|E^{(t)}\right|}\ln\left(\sqrt{\bar{\kappa}}\left(\sqrt{\bar{\kappa}} + \sqrt{\bar{\kappa}+1}\right)(n_{out}-1)\right). \end{aligned}$$

Здесь

$$\bar{\kappa} = \frac{2\omega^{(h,z)}}{\nu^{(h,z)}}$$

— среднее количество подшагов в циклах типа $\langle h, z\rangle$. При $\nu^{(h,z)} = 0$ или $\omega^{(h,z)} = 0$ или $E^{(t)} = 0$ положим, что $F(h,z,t) = 1$ и $K(h,z,t) = 0$.

2 Пример работы системы

Рассмотрим функционирование системы согласно алгоритму на примере. Пусть $n_{in} = 5$, $n_{out} = 5$, $N = 2$, а последовательность шагов работы системы приведена в Таблице 1.

t	$s_{in}^{(t)}$	$E^{(t)}$	Состояние выбора действия	$s_{out}^{(t)}$	Состояние выполнения действия
1	1	0	$\langle S1\rangle$	2	$\langle S1, A2\rangle$
2	2	0	$\langle A2, S2\rangle$	3	$\langle S2, A3\rangle$
3	3	0	$\langle A3, S3\rangle$	5	$\langle S3, A5\rangle$
4	4	1	$\langle A5, S4\rangle$	3	$\langle S4, A3\rangle$
5	3	1	$\langle A3, S3\rangle$	4	$\langle S3, A4\rangle$
6	5	2	$\langle A4, S5\rangle$	1	$\langle S5, A1\rangle$
7	1	2	$\langle A1, S1\rangle$	2	$\langle S1, A2\rangle$
8	2	2	$\langle A2, S2\rangle$	3	$\langle S2, A3\rangle$

Таблица 1: пример последовательности шагов работы системы

На Рис. 1 изображены состояния выбора действия (обведены жирной линией) и состояния выполнения действия (обведены тонкой линией), задействованные в этой последовательности шагов, и направления перехода между состояниями (жирные стрелки — переходы из состояний выбора действия). Значения возле стрелок есть приросты стимула системы. Начальным состоянием является состояние $\langle S1\rangle$, конечным — состояние $\langle S2, A3\rangle$.

Предположим, что по завершении последовательности шагов примера на шаге $t = 9$ системой был получен информационный сигнал $s_{in}^{(9)} = 3$, которому соответствует стимул $E^{(t)} = E^{(9)} = 2$, и система перешла из конечного состояния $\langle S2, A3\rangle$ в состояние выбора действия $h = \langle A3, S3\rangle$. Текущим значением $\alpha = \alpha^{(h)} = \alpha^{(\langle A3,S3\rangle)}$ является 4, поскольку последним переходом из состояния $\langle A3, S3\rangle$ был переход в состояние $\langle S3, A4\rangle$ по сигналу

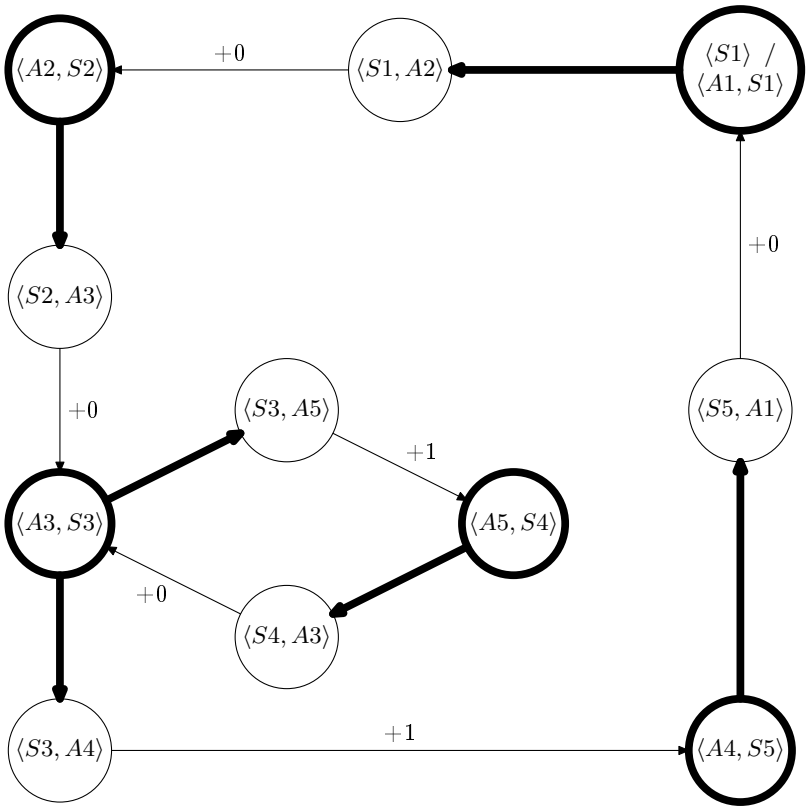

Рис. 1: пример графа переходов между состояниями системы

действия $s_{out}^{(5)} = 4$. Текущими значениями переменных $t_0^{(h)}$ и $E_0^{(h)}$, которые были занесены в момент обработки сигнала действия $s_{out}^{(5)}$, являются: $t_0^{(\langle A3,S3\rangle)} = 5$, $E_0^{(\langle A3,S3\rangle)} = E^{(5)} = 1$. Новые значения $\nu^{(h,\alpha)} = \nu^{(\langle A3,S3\rangle,4)}$, $\omega^{(h,\alpha)} = \omega^{(\langle A3,S3\rangle,4)}$ и $H^{(h,\alpha)} = H^{(\langle A3,S3\rangle,4)}$ на основе предыдущих нулевых вычисляются следующим образом:

$$\begin{aligned}
\nu^{(h,\alpha)} &:= \nu^{(\langle A3,S3\rangle,4)} + 1 = 0 + 1 = 1;\\
\omega^{(h,\alpha)} &:= \omega^{(\langle A3,S3\rangle,4)} + t - t_0^{(\langle A3,S3\rangle)} = 0 + 9 - 5 = 4;\\
H^{(h,\alpha)} &:= H^{(\langle A3,S3\rangle,4)} + E^{(9)} - E_0^{(\langle A3,S3\rangle)} = 0 + 2 - 1 = 1.
\end{aligned}$$

Так как

$$\frac{t}{E^{(t)}} = \frac{9}{2},\ \frac{H^{(\langle A3,S3\rangle,4)}}{\omega^{(\langle A3,S3\rangle,4)}} = \frac{1}{4},\ \frac{H^{(\langle A3,S3\rangle,5)}}{\omega^{(\langle A3,S3\rangle,5)}} = \frac{1}{2},\ \bar{\kappa}_4 = \frac{2\omega^{(\langle A3,S3\rangle,4)}}{\nu^{(\langle A3,S3\rangle,4)}} = 8,\ \bar{\kappa}_5 = \frac{2\omega^{(\langle A3,S3\rangle,5)}}{\nu^{(\langle A3,S3\rangle,5)}} = 4,$$

$$\begin{aligned}
\sqrt{\bar{\kappa}_4}\left(\sqrt{\bar{\kappa}_4} + \sqrt{\bar{\kappa}_4 + 1}\right)(n_{out} - 1) &= 8\left(4 + 3\sqrt{2}\right) \approx 65,9411,\\
\sqrt{\bar{\kappa}_5}\left(\sqrt{\bar{\kappa}_5} + \sqrt{\bar{\kappa}_5 + 1}\right)(n_{out} - 1) &= 8\left(2 + \sqrt{5}\right) \approx 33,8885,
\end{aligned}$$

то $F\left(\langle A3, S3\rangle, 4, 9\right) \approx 65,9411^{\frac{9}{2}\cdot\frac{1}{4}} \approx 111,314$, $F\left(\langle A3, S3\rangle, 5, 9\right) \approx 33,8885^{\frac{9}{2}\cdot\frac{1}{2}} \approx 2770,88$, $F\left(\langle A3, S3\rangle, z, 9\right) = 1$ для $z \in \{1, 2, 3\}$.

3 Композиция систем

Система ориентирована на то, что информационные сигналы, поступающие на её вход на каждом шаге взаимодействия с внешней средой, являются приближением к состояниям внешней среды. Поскольку на практике информационные сигналы обычно соответствуют не состояниям внешней среды, а её переходам из одного состояния в другое, это сильно снижает эффективность работы системы. Тогда для повышения эффективности однородная система может быть разделена на две подсистемы, первая из которых определяет текущее состояние внешней среды, а вторая на основе этого состояния генерирует оптимальное действие.

Каждый шаг взаимодействия первой подсистемы с внешней средой и второй подсистемой описывается тройкой $\left\langle \left(s_{out}^{(t-1)} - 1\right) n_{in} + s_{in}^{(t)}, \Delta E^{(t)}, s_{st}^{(t)} \right\rangle$ и заключается в приёме сигнала $\left(s_{out}^{(t-1)} - 1\right) n_{in} + s_{in}^{(t)} \in \overline{1, n_{in} \cdot n_{out}}$, составленного из сигнала действия, выданного системой на предыдущем шаге (положим, что $s_{out}^{(0)} = 1$), и информационного сигнала, полученного из внешней среды на текущем шаге, и в выдаче второй подсистеме *сигнала*

состояния внешней среды $s_{st}^{(t)} \in \overline{1, n_{st}}$. Для обеспечения логики конечного автомата, когда выдаваемый сигнал $s_{st}^{(t)}$ основывается на сигналах $s_{st}^{(t-1)}$ и $\left(s_{out}^{(t-1)} - 1\right) n_{in} + s_{in}^{(t)}$, требуется выполнение условия $N = 2$.

Каждый шаг взаимодействия второй подсистемы с первой подсистемой и внешней средой описывается тройкой $\left\langle s_{st}^{(t)}, \Delta E^{(t)}, s_{out}^{(t)} \right\rangle$ и заключается в приёме сигнала состояния внешней среды от первой подсистемы и в выдаче во внешнюю среду сигнала действия системы.

Принципом работы подсистем является рассмотренный алгоритм, в который внесён ряд изменений.

Изменениями, общими для обеих подсистем, являются изменения, направленные на повышение согласованности их работы. Пусть на некотором шаге текущим состоянием выбора действия первой подсистемы было состояние h_1, а текущим состоянием выбора действия второй подсистемы стало состояние h_2, тогда состоянием всей системы является пара $\langle h_1, h_2 \rangle$. Система может сделать цикл длины l и снова оказаться в состоянии $\langle h_1, h_2 \rangle$, если обе подсистемы выдадут те же последовательности сигналов длины l, которые обеспечили переход системы из предыдущего состояния $\langle h_1, h_2 \rangle$ в текущее состояние $\langle h_1, h_2 \rangle$. Если в процессе перехода системы из предыдущего состояния $\langle h_1, h_2 \rangle$ в текущее состояние $\langle h_1, h_2 \rangle$ первая подсистема оказывалась через каждые l_1 выданных сигналов в состоянии h_1, а вторая подсистема оказывалась через каждые l_2 выданных сигналов в состоянии h_2, то l равно наименьшему общему кратному чисел l_1 и l_2. Выдача сигналов действия подсистемами происходит из расчёта длины цикла всей системы с целью приблизительного повторения этого цикла.

Поскольку числа l_1 и l_2 могут быть дробными, то в качестве приблизительной длины цикла всей системы, не зависящей от дополнительных параметров, можно взять их произведение. Если первая подсистема находится в состоянии выбора действия h_1, и выбираемым сигналом действия является z, то при расчёте длины цикла всей системы в качестве приближения к значению l_1 можно взять среднюю длину циклов типа $\langle h_1, z \rangle$, а в качестве приближения к значению l_2 — среднюю длину цикла второй подсистемы. Если для первой подсистемы средняя длина циклов типа $\langle h_1, z \rangle$ равна κ_1 подшагов, для второй подсистемы средняя длина цикла равна κ_2 подшагов, и выбранным сигналом действия первой подсистемы является z, то, приблизительно, через $\bar{\kappa} = \kappa_1 \kappa_2$ подшагов система снова окажется в состоянии $\langle h_1, h_2 \rangle$. Аналогичный метод расчёта длины цикла всей системы используется и второй подсистемой.

Таким образом, к списку присваиваний, выполняемых при переходе в состояние выбора действия, если $2t - 1 \geq N$, добавляются увеличение количества ν_c и суммарной длины ω_c обработанных циклов (в начале работы подсистемы $\nu_c = 0$ и $\omega_c = 0$), и полный список присваиваний принимает следующий вид:

$$\begin{aligned}
\nu^{(h,\alpha)} &:= \nu^{(h,\alpha)} + 1; \\
\omega^{(h,\alpha)} &:= \omega^{(h,\alpha)} + t - t_0^{(h)}; \\
H^{(h,\alpha)} &:= H^{(h,\alpha)} + E^{(t)} - E_0^{(h)}; \\
\nu_c &:= \nu_c + 1; \\
\omega_c &:= \omega_c + t - t_0^{(h)}.
\end{aligned}$$

Для формулы относительной вероятности выбора действия (при $\nu_c > 0$) теперь необходимо взять

$$\bar{\kappa} = \frac{2\omega^{(h,z)}}{\nu^{(h,z)}} \cdot \frac{2\omega_c}{\nu_c},$$

где ω_c и ν_c относятся к другой подсистеме.

Кроме рассмотренных изменений, общих для обеих подсистем, в алгоритм работы подсистемы определения текущего состояния внешней среды внесены дополнительные изменения, направленные на увеличение произведения вероятностей появления троек $\left\langle s_{st}^{(t-1)}, \left(s_{out}^{(t-1)} - 1\right) n_{in} + s_{in}^{(t)}, s_{st}^{(t)} \right\rangle$ в истории принятых и выданных сигналов, соответствующих переходам состояний $s_{st}^{(t-1)} \to s_{st}^{(t)}$, и, следовательно, на увеличение вероятности последовательности состояний для последовательности входных сигналов.

Для этого в векторы статуса для состояний выбора действия добавлена компонента $E_1^{(h)}$, содержащая логарифм произведения вероятностей троек, встретившихся в истории принятых и выданных сигналов до шага $t_0^{(h)}$ включительно, а в векторы статистики циклов добавлена компонента $H_1^{(h,z)}$, содержащая логарифм произведения вероятностей троек, встретившихся на протяжении циклов типа $\langle h, z \rangle$, начальным значением которой является 0. Отметим, что использование логарифма величины прироста превращает прирост стимула из мультипликативного в аддитивный и, следовательно, подходящий для использования в алгоритме. Векторы статуса для состояний выбора действия принимают вид

$$\left\langle \alpha^{(h)}, t_0^{(h)}, E_0^{(h)}, E_1^{(h)}, \psi^{(h)} \right\rangle.$$

Векторы статистики циклов принимают вид

$$\left\langle \nu^{(h,z)}, \omega^{(h,z)}, H^{(h,z)}, H_1^{(h,z)} \right\rangle.$$

К списку присваиваний при переходе в состояние выбора действия, если $2t-1 \geq N$, добавлены увеличение логарифма E_m произведения вероятностей троек, встретившихся в истории принятых и выданных сигналов с момента начала работы подсистемы (когда $E_m = 0$), и увеличение $H_1^{(h,\alpha)}$ на разницу между E_m и $E_1^{(h)}$. Величина E_m увеличивается на логарифм вероятности появления тройки, вычисленной как отношение количества вхождений $\nu^{(h,\alpha)}$ тройки до шага $t_0^{(h)}$ включительно к значению $t_0^{(h)} - N/\,2$. Полный список присваиваний, выполняемых при переходе в состояние выбора действия, принимает следующий вид:

$$\begin{aligned}
\nu^{(h,\alpha)} &:= \nu^{(h,\alpha)} + 1; \\
\omega^{(h,\alpha)} &:= \omega^{(h,\alpha)} + t - t_0^{(h)}; \\
H^{(h,\alpha)} &:= H^{(h,\alpha)} + E^{(t)} - E_0^{(h)}; \\
\nu_c &:= \nu_c + 1; \\
\omega_c &:= \omega_c + t - t_0^{(h)}; \\
E_m &:= E_m + \ln \frac{2\nu^{(h,\alpha)}}{2t_0^{(h)} - N}; \\
H_1^{(h,\alpha)} &:= H_1^{(h,\alpha)} + E_m - E_1^{(h)}.
\end{aligned}$$

При использовании двух видов стимула, в качестве общей относительной вероятности выбора действия возьмём произведение относительной вероятности, вычисленной для стимула первого вида, и относительной вероятности для стимула второго вида. Руководствуясь рассуждениями, использованными при построении формулы относительной вероятности F выбора действия, введём в формулу экспоненциальный множитель, возрастающий с увеличением произведения вероятностей троек, встретившихся на протяжении циклов типа $\langle h, z\rangle$. Этот множитель переходит в слагаемое показателя степени, и формула приобретает вид

$$F_2(h,z,t) = B(h,z)^{\frac{tH^{(h,z)}}{|E^{(t)}|\,\omega^{(h,z)}}} \cdot B(h,z)^{\frac{tH_1^{(h,z)}}{|E_m|\,\omega^{(h,z)}}} = B(h,z)^{\frac{t}{\omega^{(h,z)}}\left(\frac{H^{(h,z)}}{|E^{(t)}|} + \frac{H_1^{(h,z)}}{|E_m|}\right)}.$$

После выдачи сигнала состояния внешней среды $s_{st}^{(t)}$ подсистемой, если $2t-1 \geq N$, дополнительно выполняется присваивание $E_1^{(h)} := E_m$, и полный список присваиваний принимает вид

$$\begin{aligned}
\alpha^{(h)} &:= s_{st}^{(t)}; \\
t_0^{(h)} &:= t; \\
E_0^{(h)} &:= E^{(t)}; \\
E_1^{(h)} &:= E_m.
\end{aligned}$$

4 Тестирование

Тестирование работы системы выполнялось с использованием случайных детерминированных конечных автоматов с различным количеством входных, выходных сигналов и состояний. Сигналы действия системы являлись входными для автомата. Тестирование выполнялось в режимах, когда информационными сигналами системы являлись:

1) состояния автомата, $N = 1$;

2) выходные сигналы автомата, $N = 2$;

3) выходные сигналы автомата; система состояла из подсистемы определения текущего состояния внешней среды при $N = 2$ и подсистемы генерации оптимального действия при $N = 1$.

Прирост стимула системы, равный 1, осуществлялся в случае выдачи автоматом выходного сигнала 1. Сгенерированные автоматы не содержали переходов в то же самое состояние по какому-либо входному сигналу с одновременной выдачей сигнала 1. Для сравнения со значением стимула *earned*, полученного системой в результате взаимодействия с автоматом согласно алгоритму, использовалось значение стимула *random*, полученного системой в результате случайного взаимодействия с автоматом, когда сигналы действия выбирались равновероятно из множества допустимых сигналов.

Для каждого сгенерированного автомата определялось максимальное количество стимула *maximal*, которое можно было бы получить при наиболее оптимальном взаимодействии системы с ним. Для этого в графе состояний автомата, который должен был быть связным, выполнялся поиск цикла, непрерывное повторение

которого приносит системе наибольшее количество стимула. Чтобы сделать такой поиск возможным, случайный автомат со связным графом предварительно упрощался путём замены переходов из (исходных) состояний в другие состояния переходами в исходные состояния до тех пор, пока сохранялась связность графа, и, если в результирующем автомате количество циклов превышало 100, то попытка сгенерировать подходящий для теста автомат повторялась.

Эффективность работы системы оценивалась по суммарным значениям *earned*, *random* и *maximal*, накопленным в результате 200 прогонов алгоритма, в каждом из которых использовался новый случайный автомат. Для оценки эффективности были вычислены значения

$$\begin{aligned} efr &= \frac{earned}{random} \cdot 100\%; \\ efa &= \frac{earned - random}{maximal - random} \cdot 100\%. \end{aligned}$$

Значение efr будем называть относительной, а значение efa — фактической эффективностью.

Результаты тестирования работы алгоритма для однородных систем приведены в Таблице 2, для систем, состоящих из двух подсистем — в Таблице 3. Графики зависимости efa от n_{ev} для некоторых режимов тестирования работы систем, состоящих из двух подсистем, приведены на Рис. 2-4. Каждая точка на графике соответствует значению efa, вычисленному на основе 100 прогонов алгоритма, жирная линия представляет собой регрессию.

5 Выводы

Результаты тестирования алгоритма можно рассматривать как отправную точку для сравнения различных алгоритмов взаимодействия системы с внешней средой, в качестве которой выступает детерминированный конечный автомат. Используя различные интерпретации для понятия стимула, на получение максимального количества которого нацелена система, такие алгоритмы могут быть приспособлены для решения различных задач поиска оптимального поведения. В рассмотренном алгоритме одновременно используемых видов стимула может быть несколько.

Подсистемой определения состояния внешней среды, функционирующей согласно алгоритму, решается задача приблизительного определения текущего состояния внешней среды по сигналам, переданным в среду и полученным из неё. При этом модель состояний внешней среды строится с учётом задачи получения системой максимального количества стимула из среды. К преимуществам алгоритма можно отнести то, что процесс обучения системы проходит одновременно с процессом генерации системой действий, и что каждый шаг взаимодействия с внешней средой выполняется системой за короткое фиксированное время.

Идея алгоритма может быть использована в системах управления или интерактивного взаимодействия с человеком, функционирующих в режиме реального времени.

Благодарности

Выражаю признательность своему шефу в ООО “Юникорн” (г. Киев), к. ф.-м. наук Максиму Навроцкому, за ряд ценных замечаний при критическом обзоре предварительного варианта данной статьи.

Список литературы

[1] Yoav Freund, Michael Kearns, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. Efficient Learning of Typical Finite Automata from Random Walks. *In Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, 315–324, 1993.

[2] Aaron F. Stanton, Richard E. Bleil, and Sabre Kais. A New Approach to Global Minimization. *Journal of Computational Chemistry*, Vol. 18, No. 4, 594–599, 1997.

[3] Bradley J. Clement. Learning Harmonic Progression Using Markov Models. *Genetic Algorithms in Search, Optimization, and Machine Learning*, 1998.

Кол-во состояний автомата	n_{in}	n_{out}	n_{ev}	efr, %	efa, %
Тип информационных сигналов: состояния автомата; $N = 1$					
10	10	10	5 000	849,8	77,9
10	10	10	10 000	851,5	78,2
20	20	20	5 000	1 608,6	69,3
20	20	20	10 000	1 697,4	70,9
30	30	30	15 000	2 503,5	66,9
30	30	30	30 000	2 557,1	70,1
40	40	40	35 000	3 594,1	68,7
40	40	40	70 000	3 735,6	69,7
50	50	50	65 000	4 799,3	68,7
50	50	50	130 000	4 922,0	69,1
Тип информационных сигналов: выходные сигналы автомата; $N = 2$					
10	10	10	5 000	293,9	20,2
10	10	10	10 000	313,4	22,2
20	20	20	5 000	379,4	12,8
20	20	20	10 000	415,2	14,0
30	30	30	15 000	475,3	10,4
30	30	30	30 000	519,4	12,0
40	40	40	35 000	577,7	9,4
40	40	40	70 000	607,1	9,7
50	50	50	65 000	695,6	8,7
50	50	50	130 000	704,7	8,7

Таблица 2: результаты тестирования работы алгоритма для однородных систем

Кол-во состояний автомата	n_{in}	n_{out}	n_{ev}	n_{st}	efr, %	efa, %
10	10	10	10 000	10	445,3	36,3
10	10	10	40 000	20	618,9	52,2
20	20	20	160 000	20	647,8	24,7
20	20	20	640 000	40	1 109,2	46,0
30	30	30	810 000	30	723,7	17,5
30	30	30	3 240 000	60	1 448,7	36,1
40	40	40	2 560 000	40	844,1	14,7
40	40	40	10 240 000	80	1 861,8	33,8
50	50	50	6 250 000	50	1 048,6	14,2
50	50	50	25 000 000	100	2 106,0	29,4

Таблица 3: результаты тестирования работы алгоритма для композиций систем

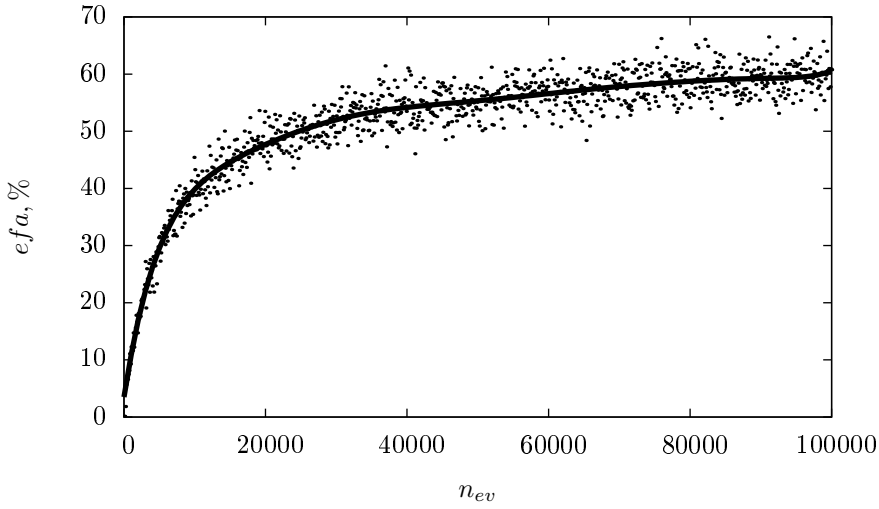

Рис. 2: автомат с 10 состояниями, $n_{in} = 10$, $n_{out} = 10$, $n_{st} = 20$

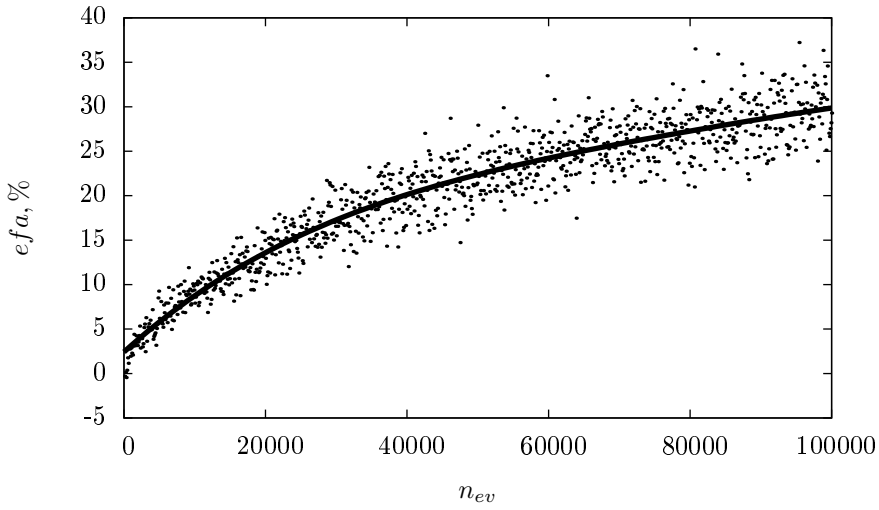

Рис. 3: автомат с 20 состояниями, $n_{in} = 20$, $n_{out} = 20$, $n_{st} = 40$

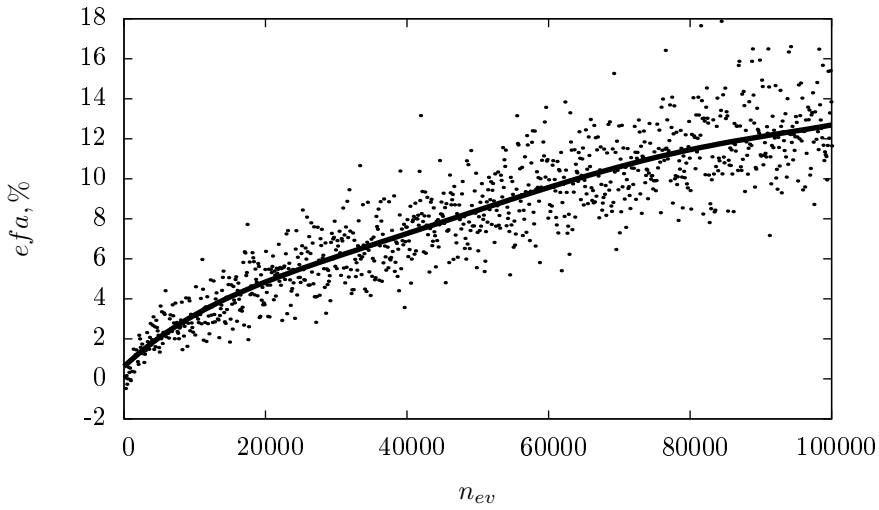

Рис. 4: автомат с 30 состояниями, $n_{in} = 30$, $n_{out} = 30$, $n_{st} = 60$