УДК 621.865.8(082)

Анализ динамики обучения робота в условиях нестационарности критериев

В.Я. Вилисов, **д.э.н., профессор кафедры Математики и естественнонаучных дисциплин,**
Государственное бюджетное образовательное учреждение
высшего профессионального образования Московской области
«Финансово–технологическая академия», г. Королев, Московская область

Представлено исследование адаптивных алгоритмов обучения оператором робототехнических систем (РТС). Оператор, как субъект целеполагания, заинтересован в том, чтобы РТС, как проводник его целей (критериев), обеспечивал максимальную эффективность их достижения. Адаптивные алгоритмы и обеспечивают адекватное отражение целей оператора в действиях РТС. Рассматриваются потенциальные возможности такой целевой адаптации РТС для класса распределительных задач.

Робот, адаптация, цель, эффективность.

Analysis of the dynamics of learning in non-stationary robot criteria

V.Ya. Vilisov, D.Sc. in economics, Professor Department of Mathematics
and Natural Sciences State,
Moscow region state–financed educational institution of higher vocational training
«Finance and technology academy», Korolev, Moscow region

The object of the research is the adaptive algorithms that are used by the operator when educating the robotic systems. Operator, being the target-setting subject, is interested in the goal that robotic systems, being the conductor of his targets (criteria), would provide a maximum effectiveness of these targets' (criteria's) achievement. Thus, the adaptive algorithms provide the adequate reflection of the operator's goals, found in the robotic systems' actions. This work considers potential possibilities of such target adaption of the robotic systems used for the class of the allocation problems.
Robot, adaptation, goal, efficiency.

Введение

Тенденция интеллектуализации роботов в последние годы становится все более очевидной. При этом развиваются как традиционное бионическое направление, воплощенное в построении аниматов [2], так и варианты обучения с учителем [3, 4, 7]. Одним из важных аспектов повышения уровня интеллекта роботов является их способность приспосабливаться (адаптироваться) к разнообразным факторам среды. Одному из направлений адаптации роботов – выявлению неявных целевых предпочтений учителя (оператора, лица, принимающего решения – ЛПР) по наблюдениям за его решениями [1] – и посвящена данная работа. При этом механизм адаптации, как и другие методы настройки параметров, требуют некоторого временного лага. Очевидно, при очень высокой динамике процессов модификации целей оператора используемая модель управления может не успевать переучиваться. Поэтому, естественно предположить, что существует некий порог динамики целей, к изменениям которых алгоритм настройки уже не успеет адаптироваться. В работе и рассматриваются предельные возможности адаптации в условиях критериальной (целевой) нестационарности работы РТС.

Постановка задачи

Распределительные задачи составляют достаточно существенную долю в числе задач, решаемых отдельным роботом или их группой [2, 3]. В качестве основного контекста будем рассматривать задачу распределения заданий или некоторого ограниченного ресурса в группе роботов. Задачи аналогичной структуры возникают и при управлении отдельным роботом, например, сервисным роботом-уборщиком или логистическими роботами, перемещающими грузы по сети дорог, или выполняющими разгрузочно-погрузочные работы на складе [3].

Одна из форм ЗЛП [5] имеет следующий вид. Целевая функция (ЦФ):

$$L(\bar{x}) = \sum_{j=1}^{n} c_j x_j. \qquad (1)$$

Критерий выбора оптимального варианта, пусть в виде оператора, максимизирующего ЦФ:

$$\bar{x}^* = arg \max_{x_j} L(\bar{x}), \qquad (2)$$

где $\bar{c} = \left\|c_j\right\|_n$ – вектор коэффициентов целевой функции; $\bar{x} = \left\|x_j\right\|_n$ – вектор варьируемых переменных; $\bar{x}^*$– оптимальное решение.

Ограничения ЗЛП часто можно представить двумя группами неравенств – одна отражает ограничения на распределяемые ресурсы:

$$A\bar{x} \le \bar{a}_0, \qquad (3)$$

где $A = \|a_{ij}\|_{mn}$ – матрица потребностей в ресурсах; $\bar{a}_0 = \|a_{i0}\|_m$ – вектор доступных к распределению ресурсов. Другая группа – ограничения на диапазон варьирования переменных:

$$\bar{x} \le \bar{b}, \bar{x} \ge 0. \qquad (4)$$

Соотношения (1) – (4) представляют собой модель выбора решений, в которой, в силу многоцелевого характера практически любой операции, выполняемой современными РТС, априорная (и текущая) неопределенность сосредоточена в векторе $\bar{c}$. Всякая новая ситуация выбора решения (управления) определяется конкретными значениями векторов $\bar{a}_0$ и $\bar{b}$, которые, как правило, измеримы и отражают состояния среды. Структура и внутренние характеристики РТС отражены в матрице A, известной и практически неизменной в течение операции.

Поскольку процедура распределения ограниченного набора ресурсов в процессе выполнения РТС операции обычно многократно повторяется (при разных ограничениях), то проблему критериальной неопределенности предлагается решить, применяя адаптивную форму ЗЛП [1] на основе решения обратной ЗЛП (ОЗЛП) и подстройки вектора $\bar{c}$ по результатам реализации выбранного решения $\bar{x}$. Оценка вектора $\bar{c}$, полученная в ходе решения ОЗЛП, является фактически аппроксимацией текущих предпочтений ЛПР, которые могут отражать множество целевых показателей, взаимосвязанных между собой по-разному и априори непредсказуемо. При этом настройка (оценивание вектора $\bar{c}$) выполняется по двухконтурной схеме, как правило, в режиме *off line* по ретроспективным данным аналогичных операций (либо в активном оптимизированном эксперименте [1]) или, если это допускает технология и динамика конкретной РТС, то в режиме *on line*. В общем случае ЦФ, аппроксимирующая предпочтения ЛПР, структурно может быть и нелинейной, например, со всеми свойствами, присущими классической функции полезности [5], тогда и алгоритмы решений обратных задач должны учитывать эту специфику.

При решении ОЗЛП используется информация о качестве принятого и реализованного решения, а также данные о системе и о текущей СТПР. Для упрощения представлений будем считать, что ограничения (4) унифицированы и добавлены к ограничениям (3). Тогда $\{A^k\}, \{\bar{a}_0^k\}, \{\bar{x}^k\}$ – последовательности наблюдений (k – номер наблюдения или цикла планирования) данных о системе и об СТПР, а $\hat{c}$ – текущий вектор оценок коэффициентов интегральной ЦФ, аппроксимирующей предпочтения ЛПР.

Алгоритм настройки

Задача построения оценок ЦФ по наблюдениям может быть решена несколькими способами [1]. Рассмотрим один из них, в котором при каждом наблюдении выполняется следующее:

1. По очередной ситуации $\bar{a}_0^k$ (СТПР) ЛПР выбирает (интуитивно или с помощью каких-то собственных механизмов) решение $\bar{x}^k$, которое ему представляется наилучшим в данной ситуации (по множеству явных и неявных показателей, принимаемых им во внимание).

2. Реализуется решение $\bar{x}^k$, а по результатам оценивается качество решения по бинарной шкале – хороший/плохой.

3. Решается обратная ЗЛП (для тех решений $\bar{x}^k$, которые признаны «хорошими»), что приводит к уточнению вектора оценок ЦФ $\hat{c}$.

При следующей СТПР цикл вновь повторяется с п.1.

Уточнение вектора $\hat{c}$ может быть выполнено, например, следующим путем [1]:

1-й этап. Для каждого k - го наблюдения выделить M активных ограничений (обычно $M = n$), для которых в точке $\bar{x}^k$ ограничения-неравенства обращаются в равенства:

$$\sum_{j=1}^{n} a_{ij}^k x_j^k - a_{i0}^k = 0. \qquad (5)$$

2-й этап. Нормировать (привести к единичной длине) все нормальные (ортогональные) векторы активных гиперплоскостей ограничений, т.е. i-е, ($i = \overline{1, M}$) векторы k-го пучка гиперплоскостей:

$$e_{ij}^k = \frac{a_{ij}^k}{\sqrt{\sum_{j=1}^{n} \left(a_{ij}^k\right)^2}} \qquad (6)$$

3-й этап. Вычислить среднюю гиперплоскость k-го пучка (т.е. суммы нормального к гиперплоскости вектора единичной длины – НВЕД) и его веса β^k. Суммарный вектор:

$$\bar{e}^k = \frac{1}{M} \sum_{i=1}^{M} \bar{e}_i^k, \qquad (7)$$

или в координатной форме:

$$e_j^k = \frac{1}{M} \sum_{i=1}^{M} e_{ij}^k, j = \overline{1, n} \qquad (8)$$

Весовой коэффициент отражает вклад наблюдения в оценку.

Чем более компактно расположены векторы в пучке наблюдений, тем более длинным будет их суммарный вектор, тогда в качестве веса наблюдения может быть использована длина суммарного вектора:

$$\beta^k = \frac{1}{M}\sqrt{\sum_{j=1}^{n}\left(\sum_{i=1}^{M} e_{ij}^k\right)^2}. \qquad (9)$$

4-й этап. Вычислить средневзвешенную гиперплоскость (средневзвешенный НВЕД) по всем K наблюдениям:

$$\bar{e}^K = \frac{1}{K}\sum_{k=1}^{K}\beta^k\bar{e}^k, \qquad (10)$$

или в координатной форме:

$$e_j^K = \frac{1}{K}\sum_{k=1}^{K}\beta^k e_j^k, j = \overline{1,n} \qquad (11)$$

Суммарный вектор $\bar{e}^K$ уже может быть использован в качестве оценки вектора ЦФ ЛПР для решения прямой ЗЛП:

$$\hat{L}^K(\bar{x}) = \sum_{j=1}^{n}\hat{c}_j^K x_j = \sum_{j=1}^{n} e_j^K x_j. \qquad (12)$$

где $\hat{c}_j^K = e_j^K;\ \hat{c}^K = [\hat{c}_1^K \quad \hat{c}_2^K \quad \cdots \quad \hat{c}_n^K]^T$.

Как показано в ряде работ автора [1, 6], настройка параметров модели предпочтений по наблюдениям выполняется достаточно быстро. Однако в ряде приложений динамика изменения обстановки, в которой действует РТС, может соизмерима с динамикой настройки параметров. В новой обстановке может изменяться иерархия ценностей ЛПР, в интересах которого действует РТС. Новая ситуация может быть обусловлена и изменениями внутри РТС, например, снижением или увеличением ее функциональности. Такие изменения и есть проявление нестационарности обстановки (обстоятельств), в которой действует РТС. Все изменения обстановки проявляются в том, что у ЛПР в новых обстоятельствах появляется новая модель предпочтений (например, представимая в виде ЦФ). Изменение параметров модели, как правило, происходит гладко, однако возможны случаи, когда изменения носят ступенчатый характер. Любой из этих видов изменений является проявлением нестационарности.

Если РТС начинает работать в стационарной среде, то обучение до приемлемого уровня эффективности функционирования происходит за некоторое время τ. В случаях проявления нестационарности, РТС должна переучиваться. И в течение времени переучивания эффективность ее работы будет отличаться от максимальной. При этом, если время переучивания РТС определяется алгоритмом обучения (находящимся «в руках» РТС-ЛПР и принимающим те или иные формы, например, пассивного оценивания в режиме нормального функционирования или активного зондирования среды и ЛПР), то динамика нестационарности может носить непредсказуемый характер. В этой связи важно знать и учитывать при планировании операций предельные возможности РТС по компенсации (с помощью переобучения) негативного воздействия целевой нестационарности.

В тех случаях, когда РТС действует в условиях активного противодействия другой стороны, то может ставиться задача создания такой нестационарности, противодействуя РТС, чтобы максимально снизить ее эффективность. Конечно, оценка предельных возможностей существенно зависит от конкретного вида и структуры РТС, а также от контекста ее использования и решаемых ею задач. С учетом ограниченного формата данной публикации продемонстрируем решение задачи на простейшем примере.

Пример

Не останавливаясь на деталях имитационных экспериментов прокомментируем основные результаты. Для двумерной ЗЛП коэффициенты истинной (имитируемой) ЦФ ЛПР в исходной позиции функционирования РТС были представлены вектором $\bar{c} = [0.8 \quad 0.6]^T$. Для рассматриваемого класса задач, без потери общности, векторы параметров ЦФ используются в нормированном виде, т.е. имеющими единичную длину, ЦФ при этом не содержит постоянную составляющую. Если РТС по наблюдениям за решениями ЛПР оценивает вектор $\bar{c}$, то время адаптации τ – это число шагов наблюдений до момента, когда решения по оценкам $\hat{c}$ с заданной доверительной вероятностью, будут совпадать с решениями, полученными по имитируемой ЦФ. Для задач различной размерности n пространства переменных это время будет разным, обозначим его τ_n. Экспериментальные (имитационные) исследования позволяют построить зависимости нормированной (т.е. приведенной к интервалу [0; 1]) средней эффективности принимаемых РТС решений в виде функции $L(t,n)$, где t – номер шага наблюдений или процесса выбора решений. Для ряда размерностей пространства решений эта функция, построенная по результатам имитационного эксперимента представлена на рисунке 1.

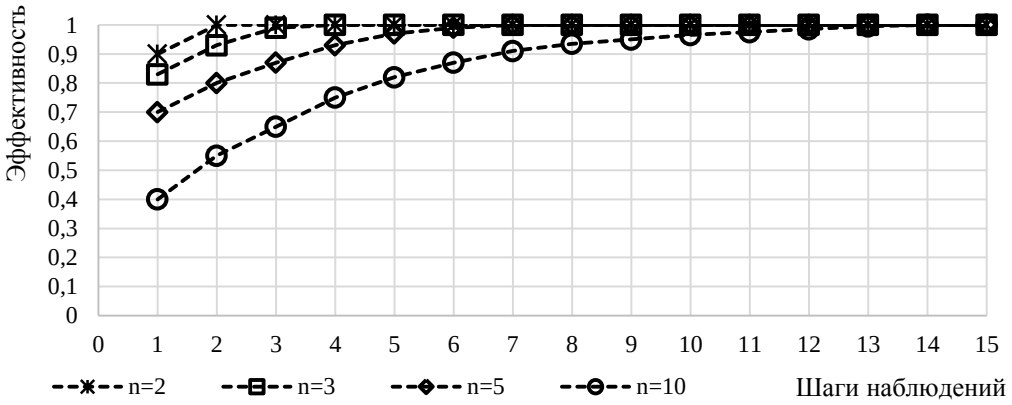

Рисунок 1 - Динамика настройки моделей различных размерностей

Нестационарность имитировалась ступенчатым изменением вектора параметров ЦФ, например, исходный вектор $\bar{c} = [0.8 \quad 0.6]^T$ скачком превращается в вектор $\bar{c} = [0.6 \quad 0.8]^T$. В момент такого скачка эффективность также скачком падает до некоторого уровня, а затем по мере обучения модели, используемой РТС для решения распределительной задачи, эффективность повышается. Таким образом, при регулярной ступенчатой смене целевых предпочтений (параметров ЦФ) график эффективности для любой размерности будет представлять собой пилообразную линию со скачкообразными падениями значения эффективности и гладкими подъемами по мере переучивания модели предпочтений РТС. При этом мерой эффективности работы РТС может служить среднее по времени значение нормированной эффективности.

Выводы

1. Эффективность использования настраиваемых по опыту ЛПР моделей распределения ресурсов в РТС существенно зависит от многочисленных параметров внешней среды функционирования, но и в не меньшей степени от адекватности представления в РТС системы предпочтений ЛПР, в интересах которого выполняются операции с помощью РТС.

2. Воздействие внешних и внутренних факторов нестационарности целевых предпочтений может быть компенсировано переобучением моделей, используемых РТС для распределительных задач управления. Однако существуют критические уровни динамики нестационарности целей, которые могут приводить к существенному снижению эффективности функционирования РТС.

Литература

1. Вилисов, В. Я. Адаптивный выбор управленческих решений. Модели исследования операций как средство хранения знаний ЛПР [Текст] / В. Я. Вилисов // Саарбрюкен (Германия).: LAP LAMBERT Academic Publishing. – 2011. – 376 с.
2. Жданов, А. А. Автономный искусственный интеллект [Текст] / А. А. Жданов // М.: БИНОМ. – 2008. – 359 с.
3. Ивченко, В. Д. Анализ методов распределения заданий в задаче управления коллективом роботов [Текст] / В. Д. Ивченко, А. А. Корнеев // Мехатроника, автоматизация, управление. – 2009. – № 7. – С. 36-42.
4. Каляев, И. А. Проблемы группового управления роботами [Текст] / И. А. Каляев, С. Г. Капустян // Мехатроника, автоматизация, управление. – 2009. – № 6. – С. 33-40.
5. Таха, Х. А. Введение в исследование операций [Текст] / Х. А. Таха // М.: Изд. дом Вильямс. – 2005. – 912 с.
6. Vilisov, V. Ya. Robot Training Under Conditions of Incomplete Information [Text] / V.Ya. Vilisov // Cornell University Library. – NY. – USA. – arXiv:1402.2996. – 14.02.2014. – Электронный ресурс. Режим доступа: http://arxiv.org/ftp/arxiv/papers/1402/1402.2996.pdf.
7. Woodward, M. P. Using Bayesian Inference to Learn High-Level Tasks from a Human Teacher [Text] / M.P. Woodward, R.J. Wood // Int. Conf. on Artificial Intelligence and Pattern Recognition, Orlando, FL, July 2009. – p. 138-145.