УДК 681.513.6

НЕЙРОСЕТЕВОЙ МЕТОД ПОДКРЕПЛЯЕМОГО ОБУЧЕНИЯ В ЗАДАЧАХ АВТОМАТИЧЕСКОГО УПРАВЛЕНИЯ

В.Н. Вичугов

Томский политехнический университет
E-mail: vlad@acs.cctpu.edu.ru

Рассмотрен метод построения адаптивных систем управления, в основе функционирования которых лежит метод обучения с подкреплением. Описано программное средство для моделирования и исследования таких систем управления. Предложен способ применения искусственных нейронных сетей для представления функции оценки воздействия.

В рамках классической теории автоматического управления при создании систем автоматического управления необходимо иметь точную математическую модель объекта управления (ОУ). Во многих реальных задачах построение такой модели либо невозможно, либо требует проведения трудоёмких исследований. При этом параметры ОУ могут изменяться в широких пределах в процессе функционирования системы, либо иметь большой разброс значений от образца к образцу. В таких случаях регуляторы с постоянными настройками не всегда могут обеспечить требуемое качество работы системы. В связи с этим актуальной является проблема построения систем автоматического управления, способных приспосабливаться к изменяющимся или неизвестным параметрам ОУ. В данной статье рассматриваются адаптивные системы автоматического управления, в основе функционирования которых лежит метод обучения с подкреплением, также называемый методом подкрепляемого обучения.

Метод подкрепляемого обучения является достаточно новым методом в группе методов машинного обучения и занимает промежуточное положение между методами обучения с учителем и без учителя. В основе метода обучения с подкреплением лежат те основополагающие принципы адаптивного поведения, которые позволяют живым организмам приспосабливаться к изменяющимся или неизвестным условиям обитания. Метод обучения с подкреплением (*Reinforcement Learning*) был представлен и подробно изложен в [1]. В данном методе в обобщенном виде рассматривается взаимодействие агента с внешней средой, в результате которого агент путем проб и ошибок самостоятельно определяет наиболее оптимальное поведение для достижения максимума некоторого критерия. Отличительной чертой метода обучения с подкреплением является наличие сигнала подкрепления, который получает агент в процессе взаимодействия с внешней средой и который является скалярной величиной, характеризующей, насколько «хорошо» функционирует агент в данный момент времени. Целью функционирования агента является максимизация суммарного сигнала подкрепления, которое получит агент при взаимодействии с внешней средой. В исходном виде метод обучения с подкреплением предполагает конечное количество состояний внешней среды и возможных воздействий агента на внешнюю среду, а также взаимодействие агента с внешней средой в дискретные моменты времени.

Указанные ограничения не позволяют свободно использовать метод обучения с подкреплением в задачах автоматического управления, т. к. сигналы в системах управления обычно являются непрерывными как по уровню, так и во времени. Тем не менее, указанный метод был успешно применен в системах управления тележкой с шестом [2], роботом, который учится плавать в водной среде [3], и перевернутым маятником [4].

На основе метода подкрепляемого обучения автором данной статьи была разработана структурная схема обобщенной системы автоматического управления, функционирующей на основе метода обучения с подкреплением (МОП-САУ), и алгоритмы работы структурных блоков. Структурная схема МОП-САУ показана на рис. 1.

Входящий в состав МОП-САУ ОУ должен удовлетворять следующим условиям:

1) ОУ является одномерным, т. е. имеет один вход и один выход;
2) в любой момент времени можно измерить вектор переменных состояния ОУ. Под переменными состояния ОУ подразумеваются сигналы, которые вместе с управляющим воздействием u однозначно определяют значение выходной величины y в будущие моменты времени.

Вектор входных сигналов устройства управления (УУ) состоит из задающего воздействия g, скорости изменения задающего воздействия g', выходной величины y и вектора переменных состояния ОУ $\boldsymbol{x}$. В результате обработки вектора входных сигналов УУ формирует управляющее воздействие u, значение которого является одним из элементов заранее определенного дискретного множества возможных воздействий A. Под действием управляющего воздействия u ОУ изменяет свое состояние.

Вектор входных сигналов поступает на вход импульсного элемента (ИЭ), который осуществляет дискретизацию по времени входных сигналов. Дискретизация по времени необходима в связи с тем, что метод обучения с подкреплением предполагает взаимодействие агента с внешней средой в дискретные моменты времени. На выходе ИЭ фор-

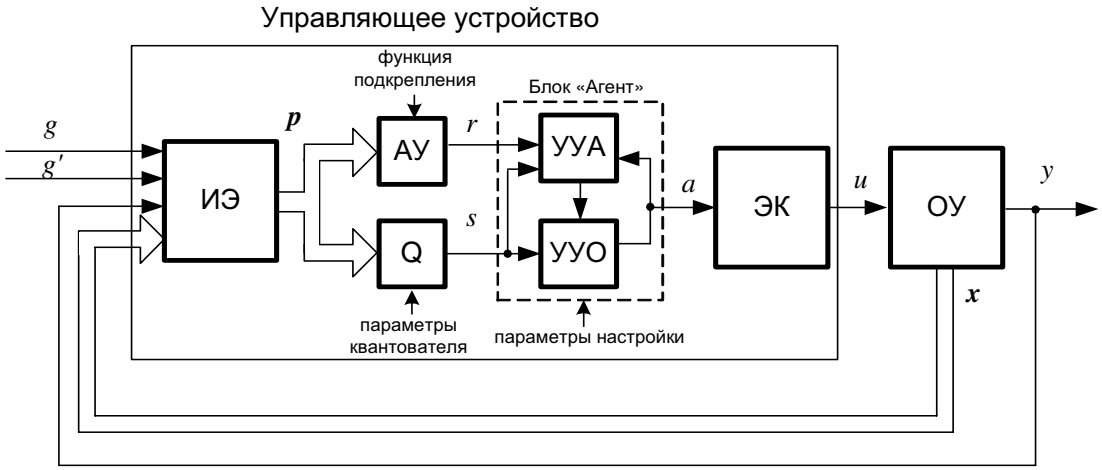

Рис. 1. *Структурная схема МОП-САУ*

мируется вектор дискретных сигналов ***p***, который поступает на анализирующее устройство (АУ) и на квантователь *Q*. АУ определяет значение сигнала подкрепления *r*, а квантователь определяет значение сигнала состояния внешней среды *s*, которое является одним из элементов заранее определенного множества возможных состояний внешней среды *S*. Экстраполятор (ЭК) переводит дискретный сигнал *a*, сформированный блоком «Агент» как воздействие на внешнюю среду, в непрерывное по времени управляющее воздействие на ОУ *u*.

Наличие в векторе входных сигналов производной входного воздействия *g*' и вектора переменных состояния ОУ ***x*** является следствием того, что в соответствии с методом обучения с подкреплением сигналы подкрепления и состояния внешней среды должны обладать свойством марковости. Несмотря на это требование, в [1] подтверждено, что метод может быть успешно применен и в том случае, когда сигналы подкрепления и состояния внешней среды не обладают свойством марковости.

Блок «Агент» является системой, функционирующей на основе метода обучения с подкреплением, и функционирует в дискретные моменты времени i=0,1,2,..., называемые тактами. В каждый момент времени i блок получает информацию о состоянии внешней среды s_i и на основе этой информации вырабатывает некоторое действие $a_i \in A(s_i)$, где $A(s_i)$ – множество действий, которые блок может выработать при текущем состоянии внешней среды s_i. В следующий дискретный момент времени i+1 блок получает оценку r_{i+1}, которая характеризует его действия на предыдущем такте, и на вход блока поступает информация о новом состоянии внешней среды s_{i+1}. Целью функционирования блока «Агент» является максимизация суммарной оценки управления [1]

$$R_i = r_{i+1} + \gamma \cdot r_{i+2} + \gamma^2 \cdot r_{i+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k \cdot r_{i+k+1},$$

где параметр $\gamma \in [0,1]$ называется параметром дисконтирования оценки управления и выбирается таким образом, чтобы величина R_i сходилась.

Для блока «Агент» внешней средой является не только ОУ, но и другие блоки УУ. Блок «Агент» состоит из двух блоков: устройства управления объектом (УУО) и устройства управления адаптацией (УУА). УУО формирует воздействие a_i на основе информации о текущем состоянии внешней среды s_i с использованием функции оценки воздействия, которая также называется *Q*-функцией [1]. УУА осуществляет коррекцию *Q*-функции на основе анализа текущего состояния внешней среды s_i и значения сигнала подкрепления r_i как результата воздействия на внешнюю среду на предыдущем такте. Эта функция имеет два аргумента: текущее состояние внешней среды s_i и некоторое воздействие *a*, которое управляющее устройство (УУ) может сформировать при s_i. Значение $Q(s_i,a)$ является суммарной оценкой управления, которую получит блок в будущем, если на текущем такте *i* сформирует воздействие *a* (т. е. a_i=*a*). Таким образом, чтобы достичь цели функционирования при точно определенной *Q*-функции и при состоянии внешней среды s_i, достаточно выбрать такой элемент *a* из множества $A(s_i)$, который соответствует максимуму функции $Q(s_i,a)$:

$$a_i = \arg\max_{a \in A} Q(s_i, a).$$

В дискретной МОП-САУ *Q*-функция представляется в виде таблицы соответствия, то есть для каждого возможного состояния внешней среды и для каждого возможного воздействия выделяется ячейка памяти, в которой хранится значение функции для данных значений аргументов. Недостатком такого варианта представления *Q*-функции является экспоненциальный рост объема требуемой памяти при увеличении количества переменных состояния ОУ, количества возможных воздействий или при увеличении количества уровней дискретизации входных сигналов УУ. В начале функционирования системы управления *Q*-функция задается произвольным образом и не содержит действительных значений суммарных оценок управления. В процессе функционирования МОП-САУ *Q*-функция корректируется, в результате её значения при-

ближаются к действительным суммарным оценкам управления. Процесс определения действительных значений Q-функции называется обучением системы. Коррекция значений Q-функции в процессе обучения осуществляется с использованием алгоритма обучения TD(λ) [1].

На основе структурной схемы и алгоритмов функционирования МОП-САУ в среде программирования *Borland Delphi* было разработано программное средство «Исследование RL-САУ», предназначенное для моделирования и исследования дискретных МОП-САУ. Программное средство позволяет задавать математическую модель ОУ в виде системы дифференциальных уравнений, определять вид и параметры задающего воздействия, задавать параметры настройки УУ, управлять процессом моделирования, отображать на экране значения всех моделируемых сигналов и их графики, определять значения показателей качества управления, сохранять результаты моделирования в файлы. Главное окно разработанной программы показано на рис. 2.

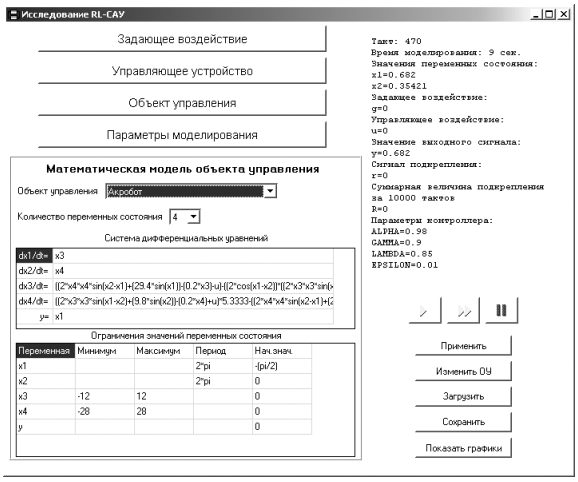

Рис. 2. *Главное окно программы «Исследование RL-САУ»*

Ниже представлены результаты экспериментальных исследований, осуществленных с помощью разработанного программного средства с дискретной адаптивной системой управления ОУ второго порядка «Колебательное звено» с передаточной функцией

$$W = \frac{1}{0,5p^2 + 0,1p + 1}.$$

На рис. 3 показана переходная характеристика данного ОУ при единичном ступенчатом воздействии. Сигнал $y(t)$ – выходной сигнал ОУ.

Для функционирования МОП-САУ необходимо установить значения параметров настройки УУ. Для проведения исследований были выбраны следующие значения параметров настройки: количество уровней квантования выходной величины – 100, первой переменной состояния – 50, второй переменной состояния – 0, задающего воздействия – 50, производной задающего воздействия – 10, возможные значения управляющего воздействия: 5, –5, 0, 15, –15. В качестве задающего воздействия был выбран прямоугольный импульсный сигнал. В эксперименте величина сигнала подкрепления равна $1-\varepsilon^2$, где ε – ошибка управления. Такое выражение было выбрано в связи с тем, что максимизация суммарной величины подкрепления приводит к минимизации величины ε. Графики, характеризующие функционирование системы в начала периода обучения, показаны на рис. 4. В УУ отсутствует априорная информация о математической модели ОУ. На рисунке видно, что управляющее воздействие в начале функционирования формируется в основном случайным образом. По мере обучения в УУ определяются точные значения Q-функции, что позволяет УУ формировать такие воздействия на ОУ, которые приведут к максимизации суммарной величины подкрепления, что приведет к минимизации среднеквадратической ошибки управления.

На рис. 5 показаны графики, характеризующие поведение системы в конце периода обучения, длительность которого составила около 7 ч модельного времени. При компьютерном моделировании на персональном компьютере среднего класса 7 ч модельного времени соответствуют около одной минуте реального. На рисунке видно, что УУ «научилось» формировать такие воздействия, которые приводят к соответствию выходного сигнала задающему сигналу. Следует учесть, что максимальная амплитуда управляющего сигнала ограничена и не позволяет добиться идеального соответствия выходного и задающего сигналов. Также следует учесть, что количество возможных значений управляющего воздействия ограничено, что не позволяет УУ установить произвольное значение этого сигнала. В конце периода обучения показатели качества управления, рассчитанные программным средством, достигли

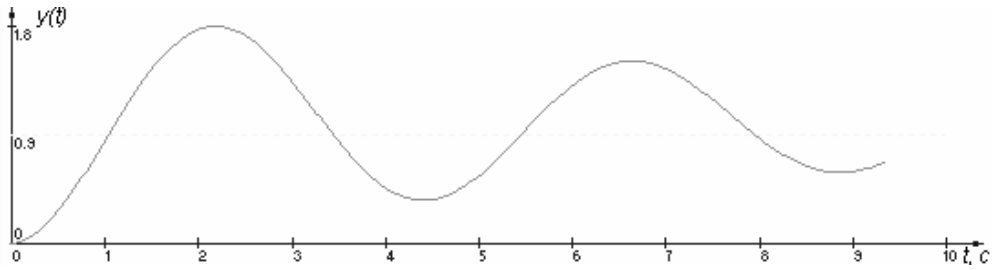

Рис. 3. *Переходная характеристика ОУ «Колебательное звено»*

следующих значений: время регулирования 0,42 с; величина перерегулирования 4,9%; среднеквадратическая ошибка управления 0,35.

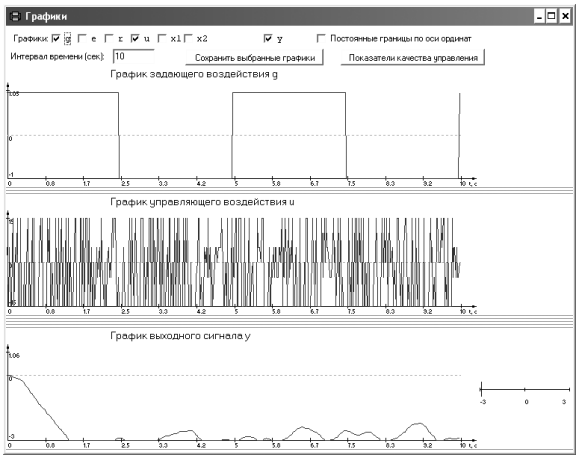

Рис. 4. *Графики задающего воздействия, управляющего воздействия и выходной величины для ОУ «Колебательное звено» в начале периода обучения*

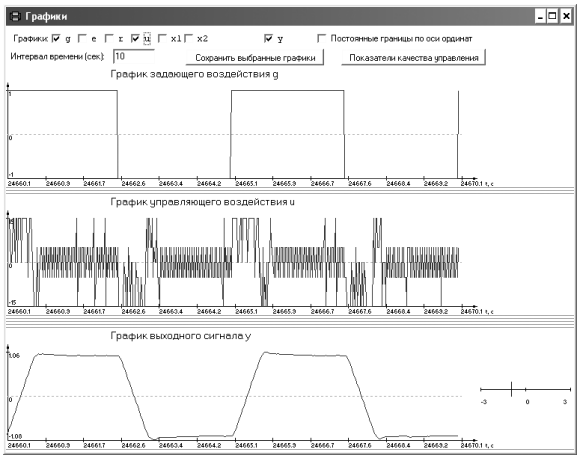

Рис. 5. *Графики задающего воздействия, управляющего воздействия и выходной величины для ОУ «Колебательное звено» в конце периода обучения*

Результаты экспериментальных исследований дискретных МОП-САУ с линейными и нелинейными ОУ второго порядка в программном средстве «Исследование RL-САУ» показали приемлемое качество управления и способность МОП-САУ адаптироваться к изменяющимся параметрам ОУ. Недостатком предложенного способа построения МОП-САУ является экспоненциальная зависимость объема требуемой памяти для представления Q-функции от порядка ОУ и от количества уровней квантования сигналов. Эту особенность исследователи в области подкрепляемого обучения называют «проклятием размерности» [1]. При математическом моделировании дискретных систем проблемы, связанные с большим объемом требуемой памяти, возникали для ОУ третьего и более высоких порядков.

Экспериментальные исследования дискретных систем с различными ОУ показали, что Q-функции являются гладкими и непрерывными, что позволяет использовать для их представления функциональные аппроксиматоры. Проблему экспоненциального роста объема требуемой памяти предлагается устранить за счет представления Q-функции на основе трехслойной искусственной нейронной сети (ИНС) прямого распространения. Так как для хранения значений параметров ИНС не требуется больших объемов памяти, их применение позволит решить указанную проблему. Кроме того, входные и выходные сигналы ИНС могут быть непрерывными, что позволяет перейти от ограниченного множества возможных состояний ОУ к непрерывному пространству состояний ОУ. Первый слой ИНС является входным и содержит столько нейронов, сколько сигналов содержится в векторе входных дискретных сигналов $\boldsymbol{p}$. Третий слой состоит из одного нейрона с линейной активационной функцией. Количество нейронов в среднем слое выбирается в зависимости от количества нейронов во входном слое. Изменение Q-функции осуществляется методом обратного распространения ошибки [5].

На рис. 6 слева показана поверхность дискретной Q-функции системы управления ОУ «Маятник», который представляет собой шест, один из концов которого прикреплён шарниром к неподвижной точке (рис. 7). Шест может свободно вращаться в вертикальной плоскости. Управляющим воздействием является вращающий момент, который вращает шест вокруг неподвижной точки. Выходной величиной объекта является угол отклонения шеста от вертикального положения θ. Целью управления является перевод маятника из исходного состояния в вертикальное положение выше оси вращения, когда угол θ равен нулю. Математическая модель ОУ представляется в виде системы дифференциальных уравнений второго порядка.

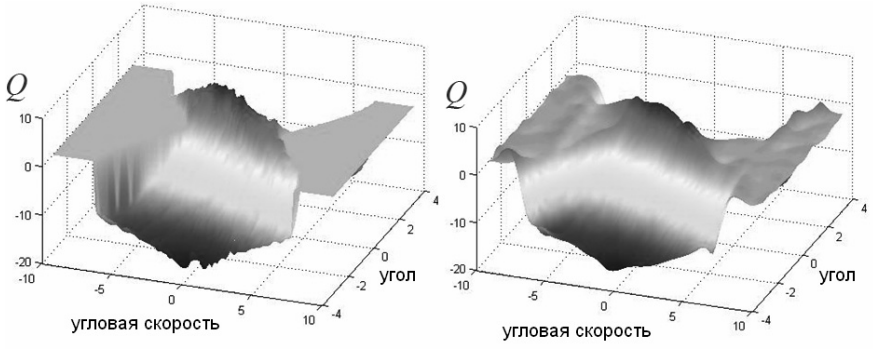

Рис. 6. *Поверхность Q-функции: дискретной (слева) и на основе ИНС (справа)*

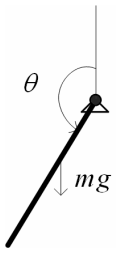

Рис. 7. *Объект управления «Маятник»*

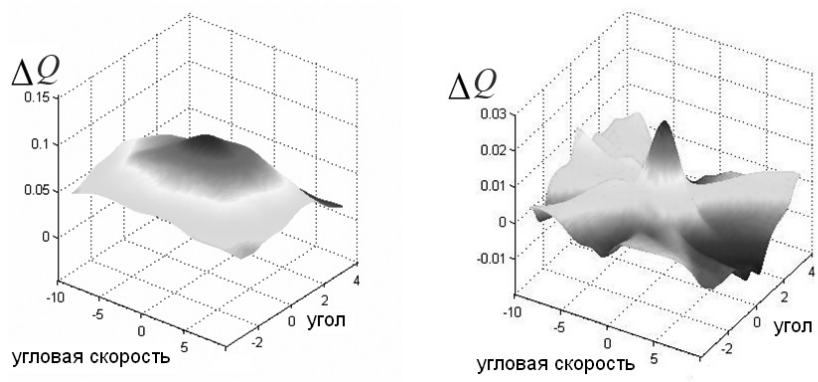

Рис. 8. *Изменение поверхности Q-функции: без закрепления окрестности изменяемой точки (слева) и с закреплением (справа)*

С помощью дискретной *Q*-функции в математическом пакете *MatLab* была обучена трехслойная ИНС, поверхность которой показана на рис. 6, справа. Среднеквадратическое отклонение значений указанных *Q*-функций друг от друга составляет 0,85, что позволяет говорить о возможности использования ИНС для представления *Q*-функций. Применение ИНС позволяет не только устранить экспоненциальную зависимость объёма требуемой памяти от порядка ОУ, но также открывает возможность создания непрерывных МОП-САУ.

Использование нейронных сетей в МОП-САУ затрудняется тем, что коррекция значения *Q*-функции на основе ИНС в одной точке приводит к изменению значений функции в других точках. На рис. 8, слева, показана поверхность изменения *Q*-функции на основе ИНС при изменении значения функции в точке (0;0) на величину приращения 0,1. На рисунке видно, что изменению подверглись все точки *Q*-функции. Это связано с тем, что применение метода обратного распространения ошибки приводит к изменению параметров связей между нейронами, которые участвуют в формировании значений функции при любых значениях входных сигналов. С целью уменьшения влияния изменения значения *Q*-функции в одной точке на значения функции в других точках был применен следующий способ обучения ИНС: совместно с изменением значения *Q*-функции в этой точке осуществляется закрепление значений *Q* функции в нескольких точках из окрестности этой точки. Закрепление осуществляется за счет применения метода обратного распространения ошибки с нулевой ошибкой. На рис. 8, справа, показана поверхность изменения *Q*-функции при решении приведенной выше задачи указанным способом. Результаты экспериментов показали, что применение такого способа итерационного обучения ИНС позволяет значительно уменьшить величину среднеквадратического отклонения значений функции в окрестности изменяемой точки от первоначальных значений. Например, для поверхностей, показанных на рис. 8, указанная величина уменьшилась с $3{,}6\cdot10^{-3}$ до $2{,}4\cdot10^{-5}$.

На основе метода обучения с подкреплением была разработана структурная схема дискретной МОП-САУ и алгоритмы функционирования структурных блоков, которые были реализованы в программном средстве «Исследование RL-САУ». Исследования линейных и нелинейных ОУ второго порядка подтвердили способность МОП-САУ достигать цели управления без априорной информации о математической модели ОУ, а также при изменении модели ОУ во время функционирования системы. В результате квантования входных сигналов уменьшается точность управления по сравнению с непрерывными системами управления, что затрудняет построение дискретных МОП-САУ для ОУ третьего и более высоких порядков. Для устранения этого недостатка предлагается представлять *Q*-функции на основе ИНС.

СПИСОК ЛИТЕРАТУРЫ

1. Sutton R.S., Barto A.G. Reinforcement learning: An introduction. – Cambridge, MA: MIT Press, 1998. – 432 p.
2. Вичугов В.Н., Цапко С.Г. Применение метода «Reinforcement Learning» в задачах автоматического управления // Современные техника и технологии: Труды XI Междунар. научно-практ. конф. студентов и молодых учёных. – Томск, 2005. – Т. 2. – С. 127–129.
3. Coulom R. Reinforcement Learning Using Neural Networks, with Applications to Motor Control. Institut National Polytechnique de Grenoble, 2002. – http://remi.coulom.free.fr/Publications/Thesis.pdf
4. Aamodt T. Intelligent Control via Reinforcement Learning. Bachelors Thesis, University of Toronto, April 1997. – http://www.eecg.utoronto.ca/~aamodt/BAScThesis/index.html
5. Антонов В.Н., Терехов В.А., Тюкин И.Ю. Адаптивное управление в технических системах. – СПб.: Изд-во С.-Петербургского ун-та, 2001. – 244 с.