УДК 519.62

А.А. ГРИШКО

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В СТОХАСТИЧЕСКИХ СИСТЕМАХ УПРАВЛЕНИЯ ЗАПАСАМИ

Рассматривается подход к решению задачи управления запасами с использованием стохастического динамического программирования и техники обучения с подкреплением, совместимой с несепарабельным критерием. В отличие от известных алгоритмов предлагаемый алгоритм обучения формирует реальное распределение прогнозируемых затрат, соответствующих рассматриваемым состояниям. Идея состоит в аппроксимации функции оценивания с помощью регистрируемых затрат. Используемая при этом стратегия базируется на робастных процедурах, гарантирующих сходимость вычислительного алгоритма. Приводится пример использования предложенного подхода для управления гидроресурсами электростанций.

1. Введение

Регулирование уровня запасов в сложных технологических комплексах зачастую определяет динамику их структуры (при этом состояние определяется уровнем запасов), а неопределенность данных (запросы, поступления, стоимость, наличие средств производства) делает структуру стохастичной. Оптимальное управление такой системой требует разработки стратегии использования запасов, позволяющей минимизировать математическое ожидание стоимости (или максимизировать математическое ожидание выигрыша) на некоторой совокупности стратегий. Таким образом, речь идет о том, чтобы знать, в какой пропорции надо использовать запасы для удовлетворения запросов и в какой степени желательно консервировать эти запасы для последующего времени [1].

Классическое решение для оптимизации выбора стратегии по сепарабельному критерию (т.е. когда стратегия может строиться шаг за шагом во времени и изменяться) может основываться на динамическом программировании (в частности, математическое ожидание является сепарабельным). К сожалению, в большинстве случаев оптимизация математического ожидания чересчур рискованная. Таким образом, целесообразно: выбрать критерий оптимизации, отличный от математического ожидания и обеспечивающий робастные решения; обеспечить возможность оптимизации такого критерия, даже если он не совместим с классической техникой динамического программирования. К классическим статистическим характеристикам случайной переменной величины C, представляющим интерес для рассматриваемой задачи, кроме математического ожидания $E(C)$ и дисперсии $\mathrm{var}(C) = E(C - E(C))$ следует отнести распространенные в системах управления рисками характеристик Risk-At-Value (RaV) и Value-At-Risk (VaR). RaV означает (для некоторого порога затрат C') вероятность того, что $C \geq C'$, а VaR для заданного порога риска соответствует минимальному C', для которого $P(C > C') \leq r$. Представления типа RaV и VaR являются наиболее полными в смысле отображения всего распределения затрат. Стохастическое динамическое программирование служит принципом классической декомпозиции для динамической оптимизации. Оно используется для оптимизации по всем сепарабельным критериям. В частности, одним из таких критериев является математическое ожидание. Однако, если принимать во внимание оценки риска VaR, то возникает проблема несепарабельности, при которой нельзя применить стандартное стохастическое динамическое программирование. Эта статья посвящена применению техники обучения с подкреплением, совместимой с несепарабельным критерием [2]. На примере обучения с подкреплением целесообразно определить стратегию, оптимизирующую компромисс «математическое ожидание/риск» несепарабельного типа $(1-\alpha)E + \alpha VaR$.

Целью данной работы является решение задачи определения стратегии управления запасами в стохастических системах с применением методов машинного обучения с подкреплением и динамического программирования.

2. Общая схема управления запасами с применением стохастической оптимизации

Общая схема стохастической динамической оптимизации и схема ее применения для управления запасами приведены на рис.1 (в левом и правом фрагментах соответственно). В самом общем ракурсе проблема стохастической динамической оптимизации может быть проиллюстрирована эволюционирующей во времени системой, которая характеризуется некоторым внутренним состоянием и пытается оптимизировать свою траекторию по заданному критерию.

Для описания такой проблемы используются:

– стохастический процесс $p(t)$ (здесь рассмотрим дискретное время от 0 до T): $p(0),...,p(T)$;

– уравнение эволюции системы: $x_{t+1} = f(x_t, u(p(t), t, x_t), p(t))$ и стратегия (регулятор) $u(\cdot,\cdot,\cdot)$, которая пытается на нее повлиять так, чтобы минимизировать математическое ожидание $c_1(x_1,u_1)+c_2(x_2,u_2)+...+c_T(x_T,u_T)$ (u_t означает $u(p(t),t,x_t)$).

Таким образом, можно сказать, что $u()$ является регулятором, минимизирующим математическое ожидание суммарных затрат.

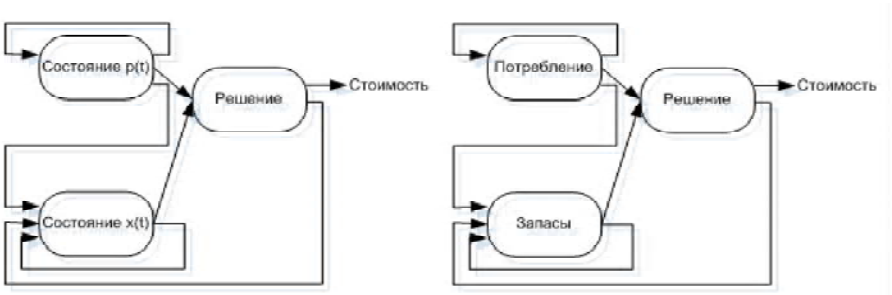

Рис. 1. Общая схема стохастической динамической оптимизации для управления запасами

В основе принципа декомпозиции Беллмана лежит способ определения $V(p(t),t,x_t)$ и, соответственно, математического ожидания $c_t(x_t,u_t)+c_{t+1}(x_{t+1},u_{t+1})+...+c_T(x_T,u_T)$ с заданными $p(t)$ и $x(t)$ для нахождения оптимальной стратегии $u()$; т.е. определение такого V, чтобы для оптимального u выполнялось следующее равенство:

$$V(\tilde{p},t,\tilde{x}) = E(c_t(x_t,u_t)+c_{t+1}(x_{t+1},u_{t+1})+...+c_T(x_T,u_T) \mid p(t)=\tilde{t}), \qquad (1)$$

где $x_t = \tilde{x}$ и $x_{t+1} = f(x_t, u(p(t),t,x_t), p(t))$.

Уравнение (1) определяет V как функцию величины, зависящей от стратегии u. Можно показать, что каждой оптимальной стратегии $u()$ соответствует своя величина $V()$. Таким образом, достаточно определить для решения задачи в каждый момент времени $u(p(t),t,x_t) \in \arg\min_u c_t(x_t,u)+V(p(t+1),t+1,x_{t+1})$. Функция $V(\cdot,\cdot,\cdot)$ называется функцией Беллмана или функцией валоризации, или же, с позиций обучения с подкреплением, критической функцией или функцией стоимости. Здесь мы будем рассматривать $V(\cdot,\cdot,\cdot)$ с одной стороны как математическое ожидание, а с другой – как компромисс вида $(1-\alpha)E+\alpha VaR$, позволяющий решать задачу оптимизации не только по матожиданию затрат, но и по риску.

Такая задача является актуальной для всех систем управления запасами, независимо от их технологического назначения. В частности, в работе [2] рассматривается управление гидравлическими запасами при использовании различных типов электростанций для удовлетворения стохастических запросов на поставку электроэнергии. При этом необходимо знать в каждый конкретный момент времени (в функции предварительно определенной модели и для текущих уровней запасов), какая часть запросов должна быть удовлетворена с помощью тепловых электростанций (ТЭЦ), а какая часть – с помощью гидроэлектрических станций (ГЭС).

Формально стохастический процесс соответствует заявкам на электроэнергию $d(t)$; продукция ТЭЦ $p_c(t)$ выбирается в соответствии с принимаемой стратегией (из диапазона $p_{min}(t) - p_{max}(t)$, определяемого эксплуатационными ограничениями теплоцентрали); продукция ГЭС $p_h(t)$ выбирается в соответствии с принимаемой стратегией (из диапазона 0–

p_h^{max}); уравнение эволюции запасов имеет следующий вид: $stock(t+1) = stock(t) + a(t) - p_h(t) - dev(t)$, где $a(t)$ – приток воды. Стратегия должна удовлетворять следующим ограничениям: $p_c(t) + p_h(t) = d(t)$, $stock(t) \geq 0$ и $stock(t) \leq stock_{max}(t)$. Функция стоимости c определяется суммой $c(t, p_c(t))$, где $c(t,\cdot)$ – функция стоимости, зависящая от t (возрастающая и выпуклая), а $c(stock(t))$ – функция штрафа за конечные запасы (убывающая). Стоимости $c_t(x_t, u_t)$ равны $c(t, p_c(t))$ для $t < T$ и $c(stock(T))$ для $t = T$. Состояние определяется парой $(d_1(t), stock(t))$ (т.е. внутренним состоянием $stock(t)$ и состоянием стохастического процесса $d_1(t)$). Стратегия $u()$ определяет $p_c(t)$ и $p_h(t)$. Динамическое программирование, используемое для поиска стратегии, работает для внутреннего состояния x_t и стохастического процесса $d_1(t)$ ($a(t)$ – детерминированная составляющая).

Рассмотрим возможность применения методов обучения с подкреплением для решения рассматриваемой задачи. Основная идея таких методов состоит в следующем: формируются модуль-актер и/или начальный модуль-критик; реализуется симуляция текущих действий модуля-актера; если используется текущий модуль-критик (алгоритмически участвующий в симуляции), модифицируются решения актера в целях оптимизации заданной функции; производится возврат к этапу симуляции.

В методах обучения с подкреплением (Q-обучение, алгоритм $TD(\lambda)$), в настоящее время широко используемых для оптимизации матожиданий, начинают применять понятия риска. В настоящей статье предлагается рассмотреть использование компромиссного критерия «матожидание»/«VaR».

Сравнивая методы обучения с подкреплением с методами классического динамического программирования, можно отметить следующее:

– классические процедуры стохастической оптимизации на основе динамического программирования обладают свойствами робастности; они реализуются с помощью алгоритмов, время вычисления для которых может быть определено заранее, а результаты вычислений являются достаточно устойчивыми;

– обучение с подкреплением хорошо подходит для работы с предысторией и, в частности, не требует структурного определения стохастического процесса, тогда как методы, основанные на стохастической оптимизации цепной предыстории с помощью динамического программирования, предполагают наличие модели;

– обучение с подкреплением позволяет естественным способом учитывать более совершенные характеристики распределения стоимости, чем сепарабельные критерии; в частности, матожидание, взвешенное риском, а не просто матожидание. Это является принципиальным моментом для данной статьи; здесь используются симуляции для представления распределения будущих стоимостей, отличные от средних будущих стоимостей. Метод, основанный только на численных переоценках (например, динамическое программирование), не может здесь работать, потому что критерий не является сепарабельным. Поэтому целесообразно разработать итеративный метод, позволяющий гарантировать сходимость функции оценивания на конечном горизонте;

– методы с подкреплением более приспособлены для работы с недискретизированными данными или с большим числом переменных состояния, когда они связаны с методами экстраполяции. Следует отметить, что объединение динамического программирования и методов, использующих экстраполяцию, все-таки менее удобно, чем обучение с подкреплением.

3. Комбинированные алгоритмы

Рассмотрим вопросы выбора схемы и алгоритмов оценивания эффективности стратегий принятия решений в системах управления запасами с использованием комбинированного критерия.

Алгоритм $TD(\lambda)$ был уже успешно проверен на практике для решения широкого круга задач; если требуется симуляция, то он хорошо сочетается с комплексными ограничениями с малыми затратами на программирование и время вычислений. Алгоритм $TD(\lambda)$ основан на использовании функции оценивания (аналога оценок Беллмана) для симуляций.

В частности, он может базироваться на симуляции различных форм величины λ (параметра, принимающего значения в диапазоне $[0,1]$). Мы используем алгоритм TD(1), который позволяет удобное введение понятия риска. В отличие от других алгоритмов TD(λ) он формирует реальное распределение будущих стоимостей, соответствующих рассматриваемым состояниям, а не будущих стоимостей, полученных по функции оценивания; оценивание «VaR» является, таким образом, прямым. Предлагаемая идея состоит в аппроксимации $V(\cdot,t,\cdot)$ с помощью регистрируемых стоимостей $c_t(x_t)+c_{t+1}(x_{t+1})+...+c_T(x_T)$. Используемая при этом стратегия является «неоптимистической», т.е. базирующейся на медленных процедурах (точнее, последовательно осуществляется некоторое количество симуляций перед началом использования функции оценивания). Такой метод является робастным по отношению к сходимости выбираемых стратегий, так как функция оценивания может сходиться без обязательной сходимости стратегии.

Введение понятия риска можно осуществить посредством использования представления типа «VaR». Модуль, позволяющий экстраполировать «VaR» для непрерывных уровней запасов на основе конечного числа примеров, предполагает необходимость введения ограничений на функции оценивания, что приводит к задаче выпуклой квадратичной оптимизации с линейными ограничениями. Структура реализации предлагаемого подхода приведена на рис. 2.

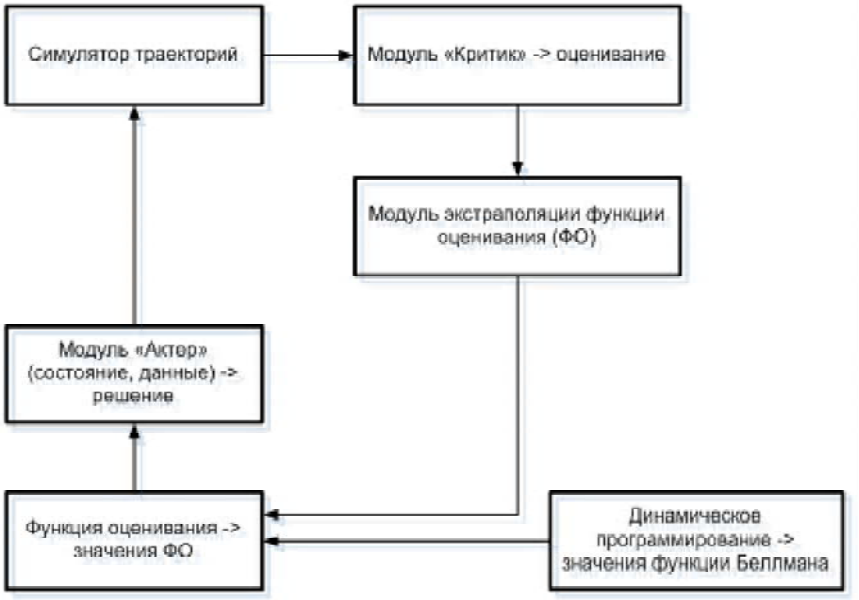

Рис. 2. Комбинированная схема формирования функции оценивания

Стохастическое динамическое программирование реализует первую функцию оценивания в случае без «VaR» и с немного упрощенным стохастическим процессом, чтобы реализовать первую функцию оценивания (т.е. начальный пункт итераций обучения с подкреплением). Комбинирование обучения с подкреплением и динамического программирования позволяет использовать гарантированную устойчивость динамического программирования и возможность методов обучения с подкреплением трансформировать оптимальное управление по матожиданию в стратегию, использующую понятие риска (после обучения с подкреплением во время фазы симуляции). Модуль экстраполяции реализует здесь экстраполяцию выпуклыми функциями, без критериев регуляризации (т.е. без штрафа на вторую производную или других штрафов этого типа). Динамическое программирование является реализуемым только для предварительных расчетов, не принимающих во внимание понятие риска; будучи теоретически факультативным, оно тем не менее обеспечивает прочную базу для операций пересчета, компенсируя, как правило, трудности параметризации обучения с подкреплением. Модуль «Критик», располагающий полной траекторией, позволяет учитывать более понятия RaV и VaR. Можно рассмотреть, например, случайную переменную C, ассоциируемую со стоимостью управления; оптимизацию $C1=(1-\alpha)E(C)+\alpha P(C>K)$, где E – символ матожидания, а P – вероятность, т.е.

учет RaV; или оптимизация $C2 = (1-\alpha)E(C) + \alpha K$, где значения K такие что $P(C > K) = 1 - \eta$, т.е. учет VaR. В случае RaV, строго говоря, надо было бы для оптимизации, например, риска превышения стоимости управления величины K, использовать функции, основанные не только на состоянии и времени, но также и на предыдущей стоимости управления. В случае VaR это не является необходимым: критерий $C2$ изменяет в функции предыдущей стоимости только одну константу и, соответственно, предыдущая стоимость не влияет на модуль «Актер».

Фиксированными являются следующие параметры: частота принятия решений (количество симуляций между каждым $V()$ и π; в соответствии с общепринятой терминологией это «степень оптимизма» метода); шаг спуска и его эволюция во времени; порог доверия применительно к понятию риска (обычно 5%); параметр компромисса между матожиданием и VaR.

Модуль «Критик» должен иметь возможность осуществлять представления в виде ломаной кривой функции оценивания $V(p,t,x)$ для заданного актера. Такая линия позволяет реализовать оптимизацию для выбора решения, соответствующую квадратичной выпуклой оптимизации с линейными ограничениями. В нашем случае p – это величина, характеризующая стохастический процесс (заявка), t – дата, x – уровень запасов. Таким образом, для каждого этапа $n = 1,2,3,...$ необходимо произвести следующие действия:

– провести симуляции (процедура симуляции и особенно дискретизации будет подробнее приведена далее);

– получить «примеры» (p,t,x,c), где t – дата, p – состояние стохастического процесса, x – уровень запасов, C – стоимость (по результатам проведенной симуляции);

– аппроксимировать матожидание будущей стоимости ломаными линиями (для шага времени и для состояния процесса: ломаная линия соединяет таким образом пары (уровень запасов, средняя будущая стоимость));

– аппроксимировать VaR ломаными линиями (для шага времени и для состояния процесса: ломаная линия соединяет таким образом пары (уровень запасов, квантиль стоимости)) и определить итоговые величины, комбинируя VaR и матожидание.

Обе аппроксимации (VaR и матожидание) получаются итеративным путем: на каждой итерации определяются симуляции VaR и матожидания для полученной стратегии управления; затем стратегия управления адаптируется для полученной функции оценивания. Пусть V_n^E и V_n^V – функции оценивания VaR и матожидания соответственно, которые были получены на n-м этапе (в виде ломаных линий). На основании V_n^E и V_n^V строятся V_{n+1}^E и V_{n+1}^V.

Пусть $V_{n+1}^E(p,t,\cdot,c)$ и $V_{n+1}^V(p,t,\cdot,c)$ – кусочно-линейные функции (с фиксированными точками изломов CSi), которые минимизируют (с указанными ниже ограничениями) сумму термов $T1, T2, T3, T4$:

$$T1 = \alpha_1 \sum_{(p,t,x,c)} (c - V_{n+1}^E(p,t,x))^2 \tag{2}$$

(терм вызова квадроплета (p,t,x,c), используемый в симуляции с применением матожидания);

$$T2 = \alpha_2 \sum_{(p,t,cs_i)} (V_n^E(p,t,cs_i) - V_{n+1}^E(p,t,cs_i))^2 \tag{3}$$

(терм вызова симуляции с предыдущего матожидания);

$$T3 = \alpha_3 \sum_{(p,t,abs_j,c)} (V_{n+1}^V(p,t,abs_j) - VaR_j)^2 , \tag{4}$$

(терм вызова оцениваемого VaR, где величины abs_j – аппроксимации VaR, полученные по данным (p,t,x,c));

$$T4 = \alpha_4 \sum_{(p,t,cs_i)} (V_n^V(p,t,cs_i) - V_{n+1}^V(p,t,cs_i))^2 , \tag{5}$$

(терм вызова предыдущей функции VaR).

Таким образом можно найти функцию, «не слишком отличную» от предыдущей функции и «не слишком далекую» от результатов симуляции. При этом принимаются во внимание следующие ограничения: $V^E_{n+1}(p,t,\cdot)$ и $V^V_{n+1}(p,t,\cdot)$ – убывающие выпуклые функции; неравенство $V^V_{n+1} \geq V^E_{n+1}$ в каждой точке (это, строго говоря, не гарантируется, но вероятность этого высока); задача не очень сложна и использование «пар» не является, следовательно, проблематичным; V_{n+1} получают в виде линейной комбинации V^E_{n+1} и V^V_{n+1}.

Выбор параметров $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ производится следующим образом: α_1 и α_3 равны 1, деленной на число термов в T1 и T3 соответственно; α_2 и α_4 равны числу игр в N уже прошедших симуляциях для рассматриваемого шага времени, деленному на число термов в T2 и T4 соответственно. Такой выбор аналогичен технике градиентного спуска; тем не менее следует понимать, что актер эволюционирует после каждого действия критика, а алгоритм градиентного спуска является разновидностью алгоритма с фиксированной точкой.

В базовом варианте предлагаемый алгоритм определяется такой последовательностью действий:

1. Определение первой функции оценивания с помощью динамического программирования.
2. Симуляции с помощью стратегии оптимизации, основанной на функции оценивания (т.е. во время симуляций получается решение, которое оптимизирует предсказанную стоимость с помощью функции оценивания).
3. Определение новой функции оценивания, которая соединяется аппроксимативно по результатам симуляции.
4. Переход к 2, в случае исчерпания лимита времени.

Формализация алгоритма зависит от параметров $1 \leq DR$ (DR – размер сигнала подкрепления), $1 \leq \Delta \leq DR$ и $0 \leq r \leq 1$, где r – уровень рассматриваемого риска:

Опишем подробнее этапы алгоритма.

Для всех T (от начального T, равного 1) с шагом, равным Δ, выполнить следующие действия: симулировать N траекторий с заданным шагом, начиная с T: начальные уровни запасов для этих траекторий, начиная с T, распределяются не поровну, так как маленькие уровни запасов являются более критичными. Пусть $L = (p_i, stock_i, pdt_i, c_i)$ – перечень квадроплетов (уровень заявок, начальный запас, шаг времени, стоимость (по результатом симуляции между шагом времени pdt и шагом времени T)), полученных в результате симуляции; L содержит $N \times DR$ пар, так как каждая симуляция позволяет получить будущую стоимость для каждого значения pdt между началом (N0) и (N0 + DR – 1).

Для всех k между 0 и DR – 1 и всех состояний p стохастического процесса между 0 и (DR – 1) в момент (k + N0) выполнить следующие действия: сформировать список L' из таких элементов L, для которых $p_i = p$ и $pdt_i = debut + k$. Пусть VaR j соответствует квантилю $1 - r$ из c_i для $stock_i$ в интервале $[a_j, a_{j+1}]$ среди элементов одного подмножества (abs_j, VaR_j) для рассматриваемых значений k и p. Поместить каждый (abs_j, VaR_j) в определенный терм $T3$. Сформировать термы $T1, T2, T4$. Оптимизировать $T1 + T2 + T3 + T4$ для получения новой функции оценивания. Симулировать большое число траекторий и вычислить полученные индикаторы риска (матожидание, кривая VaR).

4. Анализ сходимости комбинированного алгоритма

Покажем, что функция стоимости для достаточно большого N стремится к функции оценивания оптимальной стратегии, полученной по рассмотренному методу, для нормы L^2. Для этапа алгоритма, соответствующего $debut = T - 1$, сходимость соответствующей рекуррентной процедуры является очевидной. Рассмотрим гипотезу такой сходимости для этапа $debut + 1$ (с точностью $\in$). Проблема оптимизации с ограничениями по T1 (ограничения по T3 и T4 можно не рассматривать, поскольку они имеют отношение лишь к минимизации математического ожидания, а влияние $T2$ является слабым для большого

числа симуляций) сводится к асимптотическому приближению функции оценивания F к кусочно-линейной функции, соответствующей оптимальной стратегии.

Если шаг дискретизации пространства поиска выбрать в соответствии с рекомендациями, приведенными в [2], то можно получить точность $\in' + \in$ для сколь угодно малого $\in'$ при ограниченном N. При этом гарантируется, что функции оценивания после оптимизации $T1$ для $k = 0$ будут наилучшими (для нормы L^2). Отметим необходимость выполнения условия $\Delta = 1$ в случае учета рисков при реализации стратегий управления. При этом среднее значение RaV для x в интервалах $[a_j, a_{j+1}]$ является в асимптотике однозначно определяемым, так как число соответствующих дискрет по определению пропорционально отношению N/K. Поскольку функция VaR в общем случае инверсна функции RaV, то и оценки также являются сходящимися. Таким образом, для достаточно большого K (K - число сегментов $[a_j, a_{j+1}]$) выполняется условие $\min(K, 1/\max(cs_{i+1} - cs_i)) \to \infty$ (при $N \to \infty$), а следовательно, кривая значений VaR, полученная по рассмотренному алгоритму, сходится к реальной кривой VaR, построенной по результатам проведения и анализа большого числа стандартных симуляций.

5. Результаты моделирования

Моделирование рассмотренного метода проводилось для тестовой задачи управления запасами гидроресурсов электростанций с ограничениями, приведенными в [3]. Были получены кривые для комбинированного критерия $(1-\alpha)E + \alpha VaR$ при $+\alpha = 0; 0.1; 0.2; 0.4; 0.6; 0.8$ с использованием 1800 сценариев. При этом наилучшие результаты соответствуют значению $+\alpha = 0.6$. Анализ результатов показал существенное снижение уровня риска получения неудовлетворительных решений (в среднем на 55 %) при использовании комбинированного метода по сравнению с результатами, полученными с применением стандартной процедуры динамической оптимизации с сепарабельным критерием.

6. Выводы

Научная новизна полученных результатов состоит в совершенствовании методов получения оптимальных стратегий принятия решений при управлении запасами на основе применения комбинированного метода, использующего положительные свойства алгоритмов машинного обучения с подкреплением и динамической оптимизации Беллмана, который позволяет эффективно описывать динамику развития исследования стохастического процесса с учетом возможного риска и существующих ограничений.

Эффективность реализации предложенного подхода во многом зависит от правильного выбора коэффициентов в несепарабельном критерии $(1-\alpha)E + \alpha VaR$.

Практическая значимость заключается в возможности практической реализации предложенного подхода в интеллектуальных системах принятия решений с учетом возможных рисков, в частности, в системах управления запасами гидроресурсов электростанций.

Перспективным представляется развитие предложенного подхода для фиксированных уровней риска и высокого уровня неконтролируемых возмущений.

Список литературы: 1. *Dempster M. Intraday*. FX trading: An evolutionary reinforcement learning approach. Intelligent data engineering and automated learning // Proceedings of the IDEAL 2002 International Conference. 2002. P. 347-358. **2.** *Гришко А., Удовенко С., Чалая Л.* Применение гибридных методов машинного обучения в компьютерных трейдинговых системах // Системные технологии. 2010. №3(68). С. 84-92. **3.** *Sarma M., Thomas S., Shah A.* Performance Evaluation of Alternative VaR Models. Mumbai: Indira Gandhi Institute of Development Research, 2000. P. 5-10.

Поступила в редколлегию 11.12.2011

Гришко Андрей Александрович, аспирант кафедры ЭВМ ХНУРЭ. Научные интересы: методы вычислительного интеллекта, машинное обучение в системах интеллектуальной обработки данных. Адрес: Украина, 61166, Харьков, пр. Ленина, 14.