Technical Report of the 2nd Workshop on Latent Dynamics (Jun 22, 2011, Tokyo, Japan)

潜在ダイナミクスにおけるリスク考慮型意思決定

森村哲郎*
Tetsuro Morimura

Abstract: 未知の環境との相互作用のもたらすダイナミクスを解析し，意思決定を最適化する理論的枠組として強化学習がある．特に近年では，強化学習が決定的役割を果たす実問題がビジネスデータ解析や自然言語処理などの分野で次々に見出され，新しい注目が集まっている．標準的な強化学習の枠組みでは，Bellman 方程式に基づきリターン（報酬和）の期待値を推定し，意思決定を行うが，思いがけず起こる大損失のリスクの回避や，大儲けのチャンス発見のためには，リターンに関する期待値以外の情報が必要になる．ここでは，リターンの分布を推定することで，リターン分布から規定される任意の特徴量を指標とした意思決定方策を設計できることを紹介する．

Keywords: risk-sensitive decision making, reinforcement learning, return distribution

1 まえがき

潜在ダイナミクスのもと，長期リターンを最大にする戦略を探索・学習する理論的枠組みとして強化学習が知られている [1, 2, 3]．強化学習では，「何をすべきか (what)」を報酬という形で規定して，「どのように実現するか（how to）」をデータにより学習する．つまり，ダイナミクスに関する特別な知識をユーザに要請せずに，データから意思決定策を最適にすることを目指している．近年では，強化学習が決定的役割を果たす実問題がビジネスデータ解析や自然言語処理などの分野で次々に見出され，新しい注目が集まっている [4, 5, 6, 7]．

一方で，多くの強化学習法はリターンと呼ばれる報酬和の “期待値” の最大化を目的としているが [1]，期待リターンの最大化/最小化問題として定式化できない実問題も数多く指摘されている [8, 9]．例えば，起こる確率は小さいが，大きな損失が発生してしまうような可能性があり，ユーザがそのリスクをなるべく回避することに興味がある場合，期待リターンではこの目的を正しく反映しているとはいえない．つまり，期待リターンの最大化は全体としては発生するコストを軽減するであろうが，これは必ずしも高いコストの発生するリスクを積極的に回避することを目指しているわけではない．特に，金融工学において，リスク回避は主要なテーマとなっており，例えば，株式投資の場合には，小さな確率で起きる大きな損失を回避しながら収益を高めるようなポートフォリオを組むことが必要となる [10]．

このような背景から，近年，期待リターン以外のリスク指標を考慮するリスク考慮型強化学習法の研究が盛んである [11, 12, 8, 9, 13, 14, 15, 16]．特に，リターンの確率分布がわかれば，分布から規定される任意の特徴量を指標にした意思決定方策を設計が可能になるため，リターン分布推定はリスク考慮型強化学習において重要な技術になる [13, 17, 18]．

本稿では，2 節で強化学習を概説し，3 節ではリターン分布推定に関する著者らの取り組みを紹介する [19, 17, 18, 20]．これは分布 Bellman 方程式と呼ばれるリターン分布についての再帰式 [19, 18] に基づいている．

2 強化学習

2.1 節で強化学習のモデルとなるマルコフ決定過程，2.2 節で強化学習について概説する．2.3 節では，強化学習の目的関数を再考し，リスク考慮型強化学習におけるリターン分布推定の重要性を確認する．

2.1 マルコフ決定過程

強化学習のモデルとして，次の quadruplet $\{\mathcal{S}, \mathcal{A}, p_{\mathrm{T}}, P_{\mathrm{R}}\}$ で定義される離散時間マルコフ決定過程（Markov Decision Process; MDP）を考える [2, 1].

- 有限状態集合: $\mathcal{S} \ni s$,
- 有限行動集合: $\mathcal{A} \ni a$,
- 状態遷移確率分布:
  $p_{\mathrm{T}}(s|s_{-1}, a_{-1}) \triangleq \Pr(S{=}s \,|\, S_{-1}{=}s_{-1}, A_{-1}{=}a_{-1})$,

*IBM 東京基礎研究所, 242-8502 大和市下鶴間 1623-14,
e-mail tetsuro@jp.ibm.com,
IBM Research – Tokyo, 1623-14 Shimo-Tsuruma, Yamato-shi, Kanagawa 242-8502, Japan.

- 報酬観測累積分布:

$$P_{\mathrm{R}}(r|s_{-1},a_{-1},s) \triangleq \Pr(R\leq r \,|\, S_{-1}=s_{-1}, A_{-1}=a, S=s).$$

しばしば報酬 r は (s_{-1},a_{-1},s) が与えられたもとでは決定的とされるが，ここでは一般化のため，(s_{-1},a_{-1},s) が与えられても報酬は確率変数 R であるとし，その実現値を r としている．また，エージェントの行動選択確率を規定する方策には，現在の観測状態 s のみに依存するような確率的な方策族 $\Pi \ni \pi$ を考える:

$$\pi(a|s) \triangleq \Pr(A=a \,|\, S=s).$$

具体的には，以下のような状況を想定している．各時刻 t で，エージェントは方策 $\pi(a_t|s_t)$ に基づき行動 a_t を選択し，状態遷移確率 $p_{\mathrm{T}}(s_{t+1}|s_t,a_t)$ に従って次状態 s_{t+1} に遷移し，報酬確率 $P_{\mathrm{R}}(r_{t+1}|s_t,a_t,s_{t+1})$ に従って報酬 r_{t+1} を観測する．

ユーザやエージェントが調整できるものは方策 π のみである．MDP を規定する $\{\mathcal{S},\mathcal{A},p_{\mathrm{T}},P_{\mathrm{R}}\}$ は強化学習を適応する課題によって定まるものであり，一般に時間不変であり，状態遷移確率 p_{T} や報酬確率 P_{R} は未知である．

2.2 強化学習の定式化

リターン $c\in\mathbb{R}$ と呼ばれる割引報酬和（cummulative discounted reward）を定義する:

$$c \triangleq \lim_{K\to\infty}\sum_{k=1}^{K}\gamma^{k-1}r_{+k}$$

$\gamma\in[0,1)$ は減衰率と呼ばれ，問題に応じて予め設定するパラメータである．リターンは方策や状態遷移，報酬観測の確率分布に従って定まる値であるので確率変数である．ここでは，確率変数としてのリターンを C，実現値を c と書く．方策を固定とした場合，MDP はマルコフ連鎖 $\mathrm{M}(\pi)\triangleq\{\mathcal{S},\mathcal{A},p_{\mathrm{T}},P_{\mathrm{R}},\pi\}$ とみなせ，リターンの条件付き累積分布関数を

$$P_{\mathrm{C}}^{\pi}(c|s) \triangleq \Pr(C\leq c \,|\, \mathcal{S}=s, \mathrm{M}(\pi))$$

と定義し，しばしばリターン分布と呼ぶことにする．

強化学習問題は，多くの場合，リターンに関する何かしらの特徴量，特に期待値についての最大化問題と解釈できる．より具体的には，π で条件付けされる確率変数リターンについての演算子を $\mathcal{F}[c|\pi]$ と書けば，次のような最適問題として定式化できる:

$$\max_{\pi\in\Pi}\ \mathcal{F}[c|\pi]. \tag{1}$$

つまり，最適方策 $\pi^*\triangleq\operatorname*{argmax}_{\pi\in\Pi}\{\mathcal{F}[c|\pi]\}$ の探索問題であり，その目的関数は $\mathcal{F}$ である．

2.3 最適化問題としての強化学習

従来の強化学習では，目的関数 $\mathcal{F}$ に期待値が用いられ，以下が代表的である [1]:

$$\begin{aligned}\mathcal{F}[c|\pi] &\triangleq \sum_{s\in\mathcal{S}}\int_{c\in\mathbb{R}} c\; dP_{\mathrm{C}}^{\pi}(c|s)\\ &= \sum_{s\in\mathcal{S}}\mathcal{E}^{\pi}[c|s].\end{aligned} \tag{2}$$

$\mathcal{E}^{\pi}$ は $\mathrm{M}(\pi)$ で条件付けされた期待値演算子である:

$$\mathcal{E}^{\pi}[\,\cdot\,] \triangleq \mathcal{E}[\,\cdot\,|\mathrm{M}(\pi)].$$

ここでは，$\mathcal{E}^{\pi}[c|s]$ を（条件付き）期待リターンと呼ぶことにする．また，方策勾配強化学習法においては，マルコフ連鎖 $\mathrm{M}(\pi)$ は常にエルゴード性を満たすと仮定して，目的関数に

$$\begin{aligned}\mathcal{F}[c|\pi] &\triangleq \sum_{s\in\mathcal{S}} p_{\mathrm{S}}^{\pi}(s)\int_{c\in\mathbb{R}} c\, dP_{\mathrm{C}}^{\pi}(c|s)\\ &= \mathcal{E}^{\pi}[\mathcal{E}^{\pi}[c|s]]\\ &= \mathcal{E}^{\pi}[c]\end{aligned} \tag{3}$$

がしばしば用いられる [21, 1, 22]．ここで，$p_{\mathrm{S}}^{\pi}(s)$ は状態の定常分布 $\Pr(S=s\,|\,\mathrm{M}(\pi))$ である[1]．

一方で，もしユーザがリスクの制御に興味がある場合，期待リターンに基づく目的関数（式 (2) や式 (3)）では不十分である．例えば，期待リターンの最大化は全体としては発生するコストを軽減するであろうが，これは必ずしも高いコストの発生するリスクを積極的に回避することを目指しているわけではないからである [24]．また，多くの強化学習法法では最適化問題である式 (1) を解くために，式 (2) や式 (3) 内の期待値 $\mathcal{E}^{\pi}[c|s]$ を推定する必要があるが，期待値の推定は一般に頑健でないことが知られている [25]．外れ値が存在するような環境では特に問題になる．強化学習における外れ値としては，例えば，状態観測や報酬観測の失敗時の異常値などがある [26]．

つまるところ，期待リターンによる目的関数を用いた従来の強化学習法の主な問題とは，リターンについて期

[1] エルゴード性のもと，初期状態に依存しない唯一の定常分布 $p_{\mathrm{S}}^{\pi}(s)$ が存在する．また，$\mathcal{E}^{\pi}[c]$ は平均報酬 $\mathcal{E}^{\pi}[r]$ をスケール化したものと等しい [23]：

$$(1-\gamma)\mathcal{E}^{\pi}[c] = \mathcal{E}^{\pi}[r] = \lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K} r_{+k}.$$

つまり，エルゴード性のもと，$\mathcal{E}^{\pi}[c]$ を目的関数とした最適方策は，減衰率 γ によらず，$\mathcal{E}^{\pi}[r]$ を最大にする方策と等しい．

待値の情報しか見ていないことである．そこで，もしリターンについての分布推定が可能になれば，リターン分布から規定される任意の特徴量 $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_k$ 等を用いて，

$$\begin{aligned} \max_{\pi\in\Pi} \quad & \mathcal{F}_0[c|\pi], \\ \text{s.t.} \quad & \mathcal{F}_1[c|\pi] \geq \varepsilon_1, \ldots, \mathcal{F}_k[c|\pi] \geq \varepsilon_k, \end{aligned}$$

といった最適化問題を考えることが可能になる[2]．例えば，[17] では，リターンの q-分位点

$$\mathcal{Q}_q^\pi[c|s] \triangleq \inf_{c\in\mathbb{R}}\{P_{\mathrm{C}}^\pi(c|s) \geq q\}$$

に着目し，次の最適化問題を近似的に解いている:

$$\max_{\pi\in\Pi} \quad \sum_{s\in\mathcal{S}} \mathcal{Q}_q^\pi[c|s]. \tag{4}$$

q-分位点は，金融工学における主要なリスク指標である Value-at-Risk（VaR）と同義であり，ある一定の確率 $1-q$ の範囲内で起こりうる最小リターン値（もしくは最大損失額）を表すリスク指標と解釈できる．また，分位点は頑健な統計量としても知られている [28, 25]. 実際，簡単な数値実験より，式 (4) の最適化問題（の緩和問題）により得られた方策はリスク考慮型方策であり，その学習過程は頑健であったことが示されている [17].

また，あくまでも目的は期待リターンの最大化であっても，リスク指標（例えば Conditional Value-at-Risk）を利用して，積極的にリスクを負うことで，効果的な探索が達成できることも示されている [18].

以上より，リターンの分布推定技術は強化学習の新たな展開へ向けて非常に重要な要素になりえると考えられる．

3　リターン分布推定

リターン分布推定には大きく二つのアプローチがある．3.1 節で Monte Carlo 法に基づくシミュレーション・アプローチとその問題点を簡単に紹介する．3.2 節では，著者らが取り組んでいる，リターン分布の再帰方程式である分布 Bellman 方程式に用いた解析的なアプローチを紹介する．

3.1　シミュレーション・アプローチ

最も直接的なリターン分布の推定法は，Monte Carlo 法による推定であろう．つまり，各時間ステップからのリターンと状態を記憶して，時間ステップを十分進めれば，各状態からのリターン標本が多数集まるので，その標本を用いた各状態の条件付きリターン分布推定が可能となる．しかしながら，明らかに膨大なメモリーが必要であり，リターン値の確定まで（無限）時間の遅れがあるため計算コストも問題になる．そのため，Monte Carlo 法によるリターン分布推定は現実的な手法でなかった．

3.2　解析的アプローチ

リターン分布推定問題を (半) 解析的に解くための基礎となる “リターン分布についての再帰式” を 3.2.1 節で紹介する．これは，通常の期待リターンについての再帰式（Bellman 方程式）をリターン分布用に拡張したものである．3.2.2 節では，リターン分布をパーティクルにより近似し，分布 Bellman 方程式を Particle Smoothing による解く，ノンパラメトリック・リターン分布推定アルゴリズムを与える．

3.2.1　分布 Bellman 方程式

近年，期待リターンの Bellman 方程式（再帰式）を拡張した，分布 Bellman 方程式（distributional Bellman equation）と呼ばれるリターン分布の再帰式

$$\begin{aligned} P_{\mathrm{C}}^\pi(c|s) = \sum_{a,s_{+1}} & p_{\mathrm{T}}(s_{+1}|s,a)\pi(a|s) \\ & \times \int_{r_{+1}} P_{\mathrm{C}}^\pi\Big(\frac{c-r_{+1}}{\gamma}|s_{+1}\Big)\, dP_{\mathrm{R}}(r_{+1}|s,a,s_{+1}), \end{aligned} \tag{5}$$

が導出された [19, 18]. ただし，$\int_{r_{+1}} \triangleq \int_{r_{+1}\in\mathbb{R}}$, $\sum_{a,s_{+1}} \triangleq \sum_{a\in\mathcal{A}}\sum_{s_{+1}\in\mathcal{S}}$ である．また，簡便のため，式 (5) の分布 Bellman 方程式の右辺を $\mathcal{D}_\pi[c\,;\, s, P_{\mathrm{C}}^\pi]$ と書く．つまり，$\mathcal{D}_\pi$ は任意の（条件付き）累積分布 $F(c|s)$ に関する作用素

$$\begin{aligned} \mathcal{D}_\pi[c\,;\, s, F] \triangleq & \sum_{a,s_{+1}} p_{\mathrm{T}}(s_{+1}|s,a)\pi(a|s) \\ & \times \int_{r_{+1}} F\Big(\frac{c-r_{+1}}{\gamma}\,|\,s_{+1}\Big)\, dP_{\mathrm{R}}(r_{+1}|s,a,s_{+1}), \end{aligned}$$

であり，式 (5) は $P_{\mathrm{C}}^\pi(c|s) = \mathcal{D}_\pi[c\,;\, s, P_{\mathrm{C}}^\pi]$ と書ける．

分布 Bellman 方程式を解けば，その解がリターン分布である．言い換えれば，ある（累積）分布関数 $F(c|s)$ が，全ての状態 s で分布 Bellman 方程式 $F(c|s) = \mathcal{D}_\pi[c\,;\, s, F]$, $^\forall c \in \mathbb{R}$ を満たせば，F は分布 Bellman 方程式の解であり，$F = P_{\mathrm{C}}^\pi$ であることが示せる [20].

3.2.2　ノンパラメトリックなリターン分布推定法

効率良く分布 Bellman 方程式を解くアルゴリズムを導出できれば，それが効率の良いリターン分布推定法になる．しかしながら，分布 Bellman 方程式は汎関数の自

[2] 目的関数 $\mathcal{F}_0[c|\pi]$ にリターンの期待値や entropic risk measure, iterated risk measure などの時間整合性のある指標を用いないと，時間不整合性（ある時点での最適計画が，その後の時点の最適計画と一致しない）の問題が生じることが知られている [27]. また，制約 $\mathcal{F}_1, \ldots, \mathcal{F}_k$ の設定にも注意が必要である．詳しくは，[27] を参考されたい．

由度を持つため，一般に，解くことは難しい．そのため，リターン分布についてある分布族 $\mathcal{F}$ を仮定して，近似的に分布 Bellman 方程式を満たすような $F \in \mathcal{F}$ を求めるリターン分布推定法が提案されている [17, 18].

ここでは，[18] で提案された Particle Smoothing によるリターン分布推定法（Return Distribution Particle Smoothing method; RDPS）を解説する．これは，各状態もしくは各状態行動対に N 個のパーティクル

$$\boldsymbol{v}_s = \{v_{s,1}, \ldots, v_{s,N}\},\ v_{s,n} \in \mathbb{R}$$

を配置して，そのパーティクルの値のばらつきでリターン分布を近似

$$\hat{P}(c|s) \triangleq \frac{1}{N}\sum_{n=1}^{N} I(v_{s,n} \leq c) \tag{6}$$

するノンパラメトリックな分布推定法である．ここで，$I(A)$ は A が真ならば 1，偽ならば 0 を返す指示関数である．RDPS アルゴリズムのパーティクル更新手続きは，各時刻 t で，観測報酬値 r_{t+1} と一時刻先の状態 s_{t+1} のパーティクル $\boldsymbol{v}_{s_{t+1}}$ を用いて，次の手順を学習率 $\alpha \in [0, 1]$ に比例した回数繰り返すだけである:

- $n \sim \mathrm{U}(N),\ n' \sim \mathrm{U}(N)$,
- $v_{s_t,n} := r_{t+1} + \gamma v_{s_{t+1},n'}$.

ここで，:= は右辺から左変への代入演算子であり，$\mathrm{U}(N)$ は 1 から N までの自然数の一様分布である．以上より RDPS アルゴリズムは実装が非常に簡単なアルゴリズムであるが，粒子数 N を増やせば，原理的に，多峰性のあるどんな複雑なリターン分布でも推定可能である．

さらにリターン分布推定を効率化するために，RDPS アルゴリズムにエリジビレティ・トレースを適用した RDPS(λ)（λ はエリジビレティ減衰率）[18] や，分布の中心を Least square temporal difference 法 [29] で調節する方法 [20] も提案されている．また，数値実験を通して，RDPS(λ) は Monte Carlo 法よりも効率よくリターン分布推定が可能であることが示されている [18].

4 おわりに

本稿では，強化学習を概説し，リスク考慮型意思決定や強化学習の新たな展開に向けてリターンの分布推定は非常に重要であることをみた．また，著者らが行っている分布 Bellman 方程式を用いたリターン分布推定の解析的アプローチを紹介した．今後は，リターン分布推定の利用により，これまでは定式化の難しかった実問題の扱いが可能になることを期待したい．

謝辞

本稿の多くの部分は杉山将氏，八谷大岳氏（東京工業大学），鹿島久嗣氏（東京大学），田中利幸氏（京都大学）との共著論文 [18, 17, 20] に基づいている．ここに厚くお礼を申し上げる次第である．

参考文献

[1] R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, 1998.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Volumes 1 and 2*. Athena Scientific, 1995.

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[4] N. Abe, , P. Melville, C. Pendus, C. L. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk, M. Domick, and T. Gardinier. Optimizing debt collections using constrained reinforcement learning. In *International Conference on Knowledge Discovery and Data Mining*, pages 75–84, 2010.

[5] S. R. K. Branavan, H. Chen, L. S. Zettlemoyer, and R. Barzilay. Reinforcement learning for mapping instructions to actions. In *Annual Meeting of the Association for Computational Linguistics*, 2009.

[6] H. Daumé Ⅲ. *From Structured prediction to inverse reinforcement learning*. Annual Meeting of the Association for Computational Linguistics Tutorial, 2010.

[7] S. Young, M. Gašić, F. Mairesse S. Keizer, J. Schatzmann, B. Thomsona, and K. Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, 2009.

[8] O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2-3):267–290, 2002.

[9] P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under con-

straints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

[10] D. G. Luenberger. *Investment Science*. Oxford University Press, 1998.

[11] M. Heger. Consideration of risk in reinforcement learning. In *International Conference on Machine Learning*, pages 105–111, 1994.

[12] M. Sato and S. Kobayashi. Variance-penalized reinforcement learning for risk-averse asset allocation. In *Intelligent Data Engineering Automated Learning*, 2000.

[13] B. Defourny, D. Ernst, and L. Wehenkel. Risk-aware decision making and dynamic programming. In *NIPS 2008 Workshop on Model Uncertainty and Risk in RL*, 2008.

[14] H. Xu and S. Mannor. Parametric regret in uncertain markov decision processes. In *IEEE Conference on Decision and Control*, pages 3606–3613. MIT Press, 2010.

[15] E. Delage and S. Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

[16] H. Xu and S. Mannor. Distributionally robust markov decision processes. In *Advances in Neural Information Processing Systems*. MIT Press, 2010.

[17] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric return density estimation for reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

[18] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *International Conference on Machine Learning*, 2010.

[19] 中田 浩之 and 田中 利幸. マルコフ決定過程における収益分布の評価. In **情報論的学習理論ワークショップ** *(IBIS)*, 2006.

[20] 森村 哲郎, 杉山 将, 八谷 大岳, 鹿島 久嗣, and 田中 利幸. 動的計画法によるリターン分布推定. In **第** *13* **回情報論的学習理論ワークショップ**, pages 283–290, 東京, 2010.

[21] J. Baxter and P. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[22] V. S. Konda and J. N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, 2003.

[23] A. C. Singh and R. P. Rao. Optimal instrumental variable estimation for linear models with stochastic regressors using estimating functions. In *Symposium on Estimating Functions*, pages 177–192, 1996.

[24] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

[25] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[26] M. Sugiyama, H. Hachiya, H. Kashima, and T. Morimura. Least absolute policy iteration—a robust approach to value function approximation. *IEICE Transaction on Information and Systems*, E93-D(9):2555–2565, 2010.

[27] T. Osogami and T. Morimura. Time-consistency of optimization problems. In *Technical Report*. IBM Research, RT0923, 2010.

[28] A. N. Kolmogorov. The method of the median in the theory of errors. *Matermatichiskii Sbornik*, 38:47–50, 1931. Reprinted in English in *Selected Works of A.N. Kolmogorov*, vol. II, A.N. Shiryayev, (ed), Kluwer : Dordrecht.

[29] J. A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2-3):233–246, 2002.