Basis Function Discovery using Spectral Clustering and Bisimulation Metrics

(Extended Abstract)

Gheorghe Comanici
Department of Computer Science
McGill University
Montreal, QC, Canada
gcoman@cs.mcgill.ca

Doina Precup
Department of Computer Science
McGill University
Montreal, QC, Canada
dprecup@cs.mcgill.ca

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms

Algorithms, Theory, Experimentation

Keywords

Markov Decision Processes, Spectral Clustering, Basis Function Learning

1. OVERVIEW

Markov Decision Processes (MDPs) are a powerful framework for modeling sequential decision making for intelligent agents acting in stochastic environments. One of the important challenges facing such agents in practical applications is finding a suitable way to represent the state space, so that a good way of behaving can be learned efficiently. In this paper, we focus on learning a good policy when function approximation must be used to represent the value function. In this case, states are mapped into feature vectors, and a set of parameters is learned, which allows us to approximate the value of any given state. Theoretically, the quality of the approximation that can be obtained depends on the set of features. In practice, the feature set affects not only the quality of the solution obtained, but also the speed of learning.

We focus on learning feature vectors in fully specified MDPs by a set of states S, a set of actions A, a transition model $P : S \times A \times S \to [0, 1]$, and a reward function $R : S \times A \to [0, 1]$. Also, γ is a discount factor and $\gamma \in (0, 1)$. A policy $\pi : S \times A \to [0, 1]$ specifies a way of behaving for the agent, and we would like to evaluate the long term behavior it generates. We do this using the value function, which is defined (using matrix notation) as $V = \sum_{i=0}^{\infty}(\gamma\pi P)^i(\pi R) = \pi(R + \gamma P V^\pi)$. The last equality is known as the Bellman equation, and is at the heart of most incremental sampling algorithms to find V. Our goal is to linearly approximate intermediate computations of $V \approx \Phi\theta$, where Φ maps every state to feature vectors of dimension much smaller than $|S|$, and attempt to minimize $||V - \Phi\theta||_2$.

Two types of methods have been proposed in recent years to tackle this problem. The first category of methods aims to construct basis functions that reduce the error in value function estimation[3, 5]. In this case, features are reward-oriented, and are formed with the goal of reducing value function estimation errors. The second approach, exemplified by the work of Mahadevan and Maggioni [4] (and their colleagues) relies on using data to construct a state connectivity graph. Spectral clustering methods are then used to construct state features. The resulting features capture interesting transition properties of the environment (e.g. different spatial resolution) and are reward-independent. That is, the features generated are eigenvectors of the *Normalized Laplacian* [1]: $L = D_{W\mathbf{1}}^{-\frac{1}{2}}(D_{W\mathbf{1}} - W)D_{W\mathbf{1}}^{-\frac{1}{2}}$, where D_x is a diagonal matrix with entries x, and $W \in \mathcal{M}(|S|, |S|)$ is a symmetric matrix representing *diffusion models* of transitions in the underlying MDP using exploratory policies.

Our goal is to show how one can incorporate rewards in feature discovery, while still using a spectral clustering approach. We use bisimulation metrics [2], as opposed to transition information, in combination with spectral clustering.

Bisimulation Metrics are used to quantify the similarity between states in an MDP. Intuitively, states are close if their *immediate rewards* are close, and they *transition* with similar probabilities to close states. These metrics can be iteratively computed, and the number of iterations determines the accuracy of the metric. The main result of [2], and which we extend for function approximation, has usage in clustering neighboring states:

THEOREM 1: *Given a clustering map C, if V_{agg} is the value function of the aggregate MDP, then*

$||CV^*_{agg} - V^*||_\infty \leq \frac{1}{(1-\gamma)^2}||\operatorname{diag}(M^* C D^{-1}_{\mathbf{1}^T C} C^T)||_\infty$, *where M^* is the exact bisimulation metric on the original MDP.*

The above states that the approximation error is bounded above by the maximum bisimulation error between a state and the states included in the same cluster.

Eigenfunctions that incorporate reward information are desired mainly because spectral methods provide an important tool in reducing the size of representation: real positive eigenvalues corresponding to each eigenfunction. If one would have a fixed policy π, under mild conditions $\pi P = \Phi^\pi D_\lambda (\Phi^\pi)^T$ for some orthogonal Φ^π and eigenvalues λ of πP. Then $V^\pi = \Phi^\pi D_\alpha D^{-1}_{(\mathbf{1}-\gamma\lambda)}\mathbf{1}$, where $\pi R =$

Cite as: Basis Function Discovery using Spectral Clustering and Bisimulation Metrics (Extended Abstract), Gheorghe Comanici, Doina Precup, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1079-1080.

$\Phi^\pi\alpha$. Normalized Laplacian methods use an exploratory policy $\hat{\pi}$, compute an efficient alternative of $\Phi^{\hat{\pi}}$ based on W, then use as representation the eigenvectors in $\Phi^{\hat{\pi}}$ with high-order $1/(1-\gamma\lambda)$. As noticed, D_α, the representation of the reward using the proposed features, is completely ignored, and bisimulation metrics are going to provide alternatives to $\Phi^{\hat{\pi}}D_\alpha$, by combining reward and transition information to generate measures of similarity.

Extending bisimulation bounds for general feature maps: The main extension that allows one to use bisimulation as a heuristic for feature generation is that feature sets that are faithful to the bisimulation metric provide better bounds on the approximation error.

Given a feature extractor with the property $Q\mathbf{1} = \mathbf{1}$, we compute the optimal value function V^*_ϕ of the induced MDP with on the feature set: $P_\Phi = D^{-1}_{\Phi^T\mathbf{1}}\Phi^T P\Phi$ and $R_\Phi = D^{-1}_{\Phi^T\mathbf{1}}\Phi^T R$. This can than be used to obtain the largest representable value function as ΦV^*_ϕ. The following theorem generalizes previous results on clustering:

THEOREM 2: *Given an MDP, let Φ be a set of feature vectors with the property $\Phi\mathbf{1} = \mathbf{1}$. Then the following holds:*

$$||\Phi V^*_\Phi - V^*||_\infty \leq ||\operatorname{diag}(M^*\Phi D^{-1}_{\mathbf{1}^T\Phi}\Phi^T)||_\infty/(1-\gamma)^2$$

2. EMPIRICAL RESULTS

One is free to use any kind of feature selections, but if these impose a relationship faithful to the bisimulation metric, then one has theoretical guarantees that the error in approximation is bounded. To illustrate this, we modify the spectral decomposition methods presented in [4] to use the bisimulation metric. In this end, we use a similarity matrix W_K, which is the inverse exponential of M^*, normalized in $[0,1]$. We compare it to previous methods based solely on state-topology (i.e. $W_T(s,s') = 1$ if and only if one can transition $s \to s'$ or $s' \to s$).

We first compute the eigenvectors of $D^{-\frac{1}{2}}_{W\mathbf{1}}(D_{W\mathbf{1}} - W)D^{-\frac{1}{2}}_{W\mathbf{1}}$, where W is either of W_K or W_T. We select the first k eigenvectors of F, based on the corresponding eigenvalues. The exact value of V^π is then computed as $(I-\gamma\pi P)^{-1}\pi R$, and then compared to ΦV^Φ. The later is simply V^π's projection on an orthonormal basis of Φ, which in turn is an application of the Gram-Schmidt procedure.

7x7 and 9x11 grid worlds (Figure 1) are controlled by 4 actions representing the four movement directions in a grid. Upon using any action, the corresponding movement is performed with probability 0.9, and the state does not change with probability 0.1. If the corresponding action results in collision with wall, the state does not change. Rewards of 10 are obtained upon entering goal states (labelled by dots).

Empirical Results are shown in Figure 2 as comparisons between the best approximations possible using variable number of features. For a number of 300 randomly generated policies, the presented method was used to compute the best approximation to the value function using both bisimulation and the accessibility matrix for state similarity (as previously presented in Mahadevan and Maggioni [4]). The graphs represent average L_2-error in approximation. The last two graphs were generated by running the same algorithm at different numerical precision of the bisimulation metric.

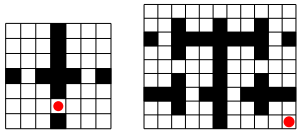

Figure 1: 7x7 and 9x11 Grid Worlds

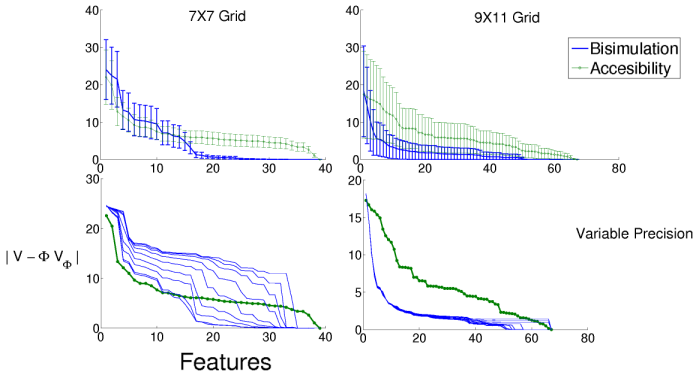

Figure 2: Empirical Results

3. CONCLUSION AND FUTURE WORK

We presented an approach to automatic feature construction in MDPs based on using bisimulation metrics and spectral clustering. The main aspect of this work is that we obtain features that are *reward-sensitive*, which proves quite important in practice, according to our experiments. Even when the precision of the metric is reduced, to make computation faster, the features we obtain still allow for a very good approximation. The use of bisimulation allows us to obtain solid theoretical guarantees on the approximation error. These are obtained by extending previous results on clustering using bisimulation to more general function approximation settings. However, the cost of computing or even approximate bisimulation metrics may be prohibitive for some domains. The results presented here are meant as a proof-of-concept to illustrate the utility of bisimulation metrics for feature construction. We are currently exploring more efficient reward-based feature construction methods.

Acknowledgements : This work was founded in part by FQRNT and ONR. We also want to thank the anonymous reviewers for their useful comments.

4. REFERENCES

[1] F. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, 1997.

[2] N. Ferns, P. Panangaden, and D. Precup. Metrics for Finite Markov Decision Processes. In *NIPS*, 2003.

[3] P. W. Keller, S. Mannor, and D. Precup. Automatic Basis Function Construction for Approximate Dynamic Programming and Reinforcement Learning. In *ICML*, 2006.

[4] S. Mahadevan and M. Maggioni. Proto-value functions: A Laplacian Framework for Learning Representation and Control in Markov Decission Processes. *Machine Learning*, 2005.

[5] R. Parr, H. Painter-Wakefiled, L. Li, and M. L. Littman. Analyzing Feature Generation for Value Function Approximation. In *ICML*, 2007.