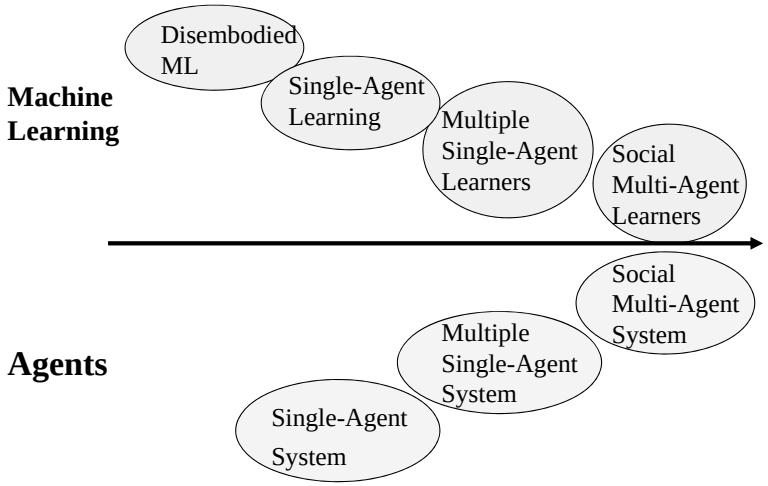
Reinforcement learning in Multi-Agent Systems

Learning in Multi-Agent Systems: Important Issues

- Classification
- Social Awareness
- Communication
- Role Learning
- Distributed Learning
- Focus: Learning of Coordination

A Brief History



Types of Multi-Agent Learning [Weiss & Dillenbourg 99]

- **Multiplied Learning:** No interference in the learning process by other agents (except for exchange of training data or outputs).
- **Divided Learning:** Division of learning task on functional level.
- **Interacting Learning:** cooperation beyond the pure exchange of data.

Social Awareness

- Awareness of existence of other agents and (eventually) knowledge about their behavior.
- Not necessary to achieve near optimal MAS behavior: rock sample collection [Steels 89].
- Can it degrade performance?

Levels of Social Awareness [Vidal&Durfee 97]

- **0-level agent:** no knowledge about existence of other agents.
- **1-level agent:** recognizes that other agents exist, model other agents as 0-level.
- **2-level agent:** has some knowledge about behavior of other agents and their behavior; model other agents as 1-level agents.
- **k-level agent:** model other agents as (k-1)-level.

Social Awareness and Q Learning

- 0-level agents already learn *implicitly* about other agents.
- [Mundhe and Sen, 00]: study of two Q learning agents up to level 2.
- Two 1-level agents display slowest and least effective learning (worse than two 0-level agents).

Agent models and Q Learning

- $Q: S \times A^n \rightarrow R$, where n is the number of agents.
- If other agent's actions are not observable, need assumption for actions of other agents.
- **Pessimistic assumption:** given an agent's action choice other agents will minimize reward.
- **Optimistic assumption:** other agents will maximize reward.

Agent Models and Q Learning

- Pessimistic Assumption leads to overly cautious behavior.
- Optimistic Assumption guarantees convergence towards optimum [Lauer & Riedmiller '00].
- If knowledge of other agent's behavior available, Q value update can be based on probabilistic computation [Claus and Boutilier '98]. *But:* no guarantee of optimality.

Q Learning & Communication [Tan 93]

Types of communication:

- Sharing sensation
- Sharing or merging policies
- Sharing episodes

Results:

- Communication generally helps
- Extra sensory information may hurt

Role Learning

- Often useful for agents to specialize in specific roles for joint tasks.
- Pre-defined roles: reduce flexibility, often not easy to define optimal distribution, may be expensive.
- How to learn roles?
- [Prasad et al. 96]: learn optimal distribution of pre-defined roles.

Q Learning of roles

- [Crites&Barto 98]: elevator domain; regular Q learning; no specialization achieved (but highly efficient behavior).
- [Ono&Fukumoto 96]: Hunter-Prey domain, specialization achieved with *greatest mass merging strategy*.

Q Learning of Roles [Balch 99]

- Two main types of reward function: local and global.
- Global reward supports specialization.
- Local reward supports emergence of homogeneous behaviors.
- Some domains benefit from learning team heterogeneity (e.g., robotic soccer), others do not (e.g., multi-robot foraging).
- Heterogeneity measure: social entropy.

Distributed Learning

- Motivation: Agents learning a global hypothesis from local observations.
- Application of MAS techniques to (inductive) learning.
- Applications: Distributed Data Mining [Provost & Kolluri '99], Robotic Soccer.

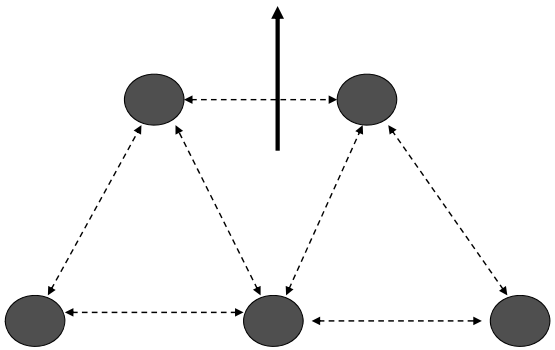
Distributed Data Mining

- [Provost & Hennessy 96]: Individual learners see only subset of all training examples and compute a set of local rules based on these.
- Local rules are evaluated by other learners based on their data.
- Only rules with good evaluation are carried over to the global hypothesis.

Learning to Coordinate

- Good coordination is crucial for good MAS performance.
- Example: soccer team.
- Pre-defined coordination protocols are often difficult to define in advance.
- Needed: learning of coordination.
- Focus: Q-learning of coordination.

Soccer Formation



Soccer Formation Control

- Formation control is a coordination problem.
- Good formations and set-plays seem to be a strong factor in winning teams.
- To date: pre-defined.
- Can (near-)optimal formations be (reinforcement) learned?

A Sub-Problem

- Given: *n* agents at random positions, and a formation having *n* positions.
- Wanted: set of *n* policies that transforms initial state into the desired formation.
- Specifically: Q learning of these policies.

A Further Simplification

- MAS Policy: decision procedure who takes which position.
- No two agents should choose the same formation position.
- Problem reduces to reinforcement learning of coordination in cooperative games.

Cooperative Games

- Players perform actions simultaneously.
- Afterwards, all players receive the same reward based on the joint action.

		Player 2	
		A1	A2
Player 1	A1	5	3
	A2	2	0

Cooperative Games and Formations

- Consider 2-player formation with 2 positions: left, right.
- Corresponding cooperative game:

		Player 2	
		left	right
Player 1	left	0	5
	right	5	0

Learning in Cooperative Games

- To date: focus on Q-learning.
- Is communication/observation amongst agents necessary?
- Does this requirement change with increasing difficulty of the cooperative game?

Convergence

- Single-agent Q-learning: guaranteed convergence (to optimum).
- Multi-agent Q-learning: more assumptions needed.
- Crucial in MAS: action selection strategy.

Q Learning Revisited

- Modified Q update function:
  $Q(a) = Q(a) + \gamma\,(r - Q(a))$
- Boltzmann action selection strategy:

$$P(a) = \frac{e^{EV(a)/T}}{\sum_{a'} e^{EV(a')/T}}$$

Boltzmann Exploration

- Usually: EV(a) = Q(a).
- Trade-off between *exploration* and *exploitation*.
- Higher temperature T results in more emphasis on exploration.
- Temperature T should be high at first, and lowered with time ($T(t) = e^{(-s*t)}$).

Q Learning of Coordination

- [Singh et al., 2000]: convergence to *some* joint action can be ensured with specific temperature properties.
- Convergence to optimal joint action for simple cases:

		Player 2	
		A1	A2
Player 1	A1	5	3
	A2	2	0

"Difficult" Cooperative Games

- Climbing Game [Claus & Boutillier, 98]:

		Player 2		
		a	b	c
Player 1	a	**11**	-30	0
	b	-30	7	6
	c	0	0	5

Climbing Game

- Multiplied Q learning with Boltzmann exploration converges to suboptimal (c,c).
- [C & B, 98]: Joint action learners (JAL).
- Agents observe each others actions and build a probabilistic model, according to which the next action is chosen.
- Agents get to (b,b) but are stuck there.

Climbing Game (cont.)

- Optimistic assumption [Lauer & Riedmiller, 00]: never reduce Q-values due to penalties.
- Converges quickly to optimal (a,a).
- However, does not converge on stochastic version of climbing game.

Stochastic Climbing Game

		Player 2		
		a	b	c
Player 1	a	**12 / 10**	0/-60	0/-60
	b	0/-60	14/0	8/4
	c	5/-5	5/-5	7/3

FMQ Heuristic

- [Kapetanakis & Kudenko, 02]:
  - $EV(a) = Q(a) + c\ freq(maxR(a))\ maxR(a)$
- EV(a) carries information on how frequently an action produces its maximum corresponding reward.
- Converges to optimal (a,a) for climbing game and *partially stochastic* climbing game.

Partially Stochastic Climbing Game

Player 1 \ Player 2	a	b	c
a	**11**	0/-60	0/-60
b	0/-60	14/0	8/4
c	5/-5	5/-5	7/3

"Difficult" Cooperative Games

- Penalty Game [Claus & Boutillier, 98]

Player 1 \ Player 2	a	b	c
a	**10**	0	k
b	0	2	0
c	k	0	**10**

Penalty Game

- JAL: convergence to optimal (a,a) or (c,c) only for small penalties k (k>-20).
- Both optimistic assumption and FMQ converge to either optimum also for large penalties (up to –100).

Learning of Coordination: More Questions

- Scaling-up of Q learning approaches?
- Agents with state: [Boutillier, 99].
- Large numbers of actions/agents?
- Learning of formations from non-explicit rewards?

Learning of Coordination: Conclusions

- Idealized and simple cases have been studied and solved.
- Mutual communication/observation may not be needed.
- Beyond Q learning: Evolutionary approaches [Quinn, 01].