報酬に基づいた環境情報の取捨選択による行動学習の効率化に関する研究

氏　　名	木島　康隆（きしま　やすたか）	
学位論文題目	報酬に基づいた環境情報の取捨選択による行動学習の効率化に関する研究	
論文審査委員	主　査	教　授　佐賀　聡人
		教　授　畑中　雅彦
		准教授　須藤　秀紹
		准教授　高氏　秀則

論文内容の要旨

本論文では，強化学習における学習の効率化に関して，ロボット外部の情報とロボットの内部の情報の２つの情報から考察する。強化学習では，Q 空間と呼ばれる状態軸，行動軸，Q 値軸からなる学習空間を基に学習を行う。状態軸はロボットが観測した周囲の環境の状態を示す。行動軸はロボットがとることの出来る行動を示す。Q 値軸はある状態である行動をとった時に得られる期待報酬値を示す。Q 空間は報酬を基に更新される。

強化学習の問題点として，学習に時間がかかるという問題が挙げられる。特に，ロボットに搭載されるセンサが増加し，環境状態の情報量が増えると，それに伴い状態軸も増大し学習空間が大きくなる。学習空間が大きくなるとそれだけ多くの経験を必要とする。その結果，学習に多くの時間を要する。

この問題に対して，本研究ではロボットの外部と内部の情報から Q 値を改変し学習を効率化させる手法を提案する。ロボットの外部の情報とは，他のロボットとのコミュニケーションによって得る他のロボットの経験情報（Q 値）である。実社会では，時間的な制約によりロボットが獲得可能な情報には限りがある。そのため，他のロボットとのコミュニケーションにより自身が得た Q 値に加え他者からの Q 値により，学習をより効率的にすることを考える。しかし，ただコミュニケーションを行うだけでは，自身の学習を阻害するような情報を得てしまい，却って学習の効率を下げる恐れがある。コミュニケーションを行う相手に関して選別し，自身にとって有益な情報をもたらす他者とコミュニケーションすべきである。そこで，本研究では，自身とって有益な情報を持つ他者を基に学習しコミュニケーションする

ことで，効率的に学習を行う手法を提案する。次に，ロボットの内部の情報の取り扱いとして，Q 空間そのものをタスクに適した形に改変する。タスクを遂行するにあたり，センサ情報全てが必要であるとは限らない。タスクによって，重要となるセンサ情報と不要なセンサ情報が存在する。ロボットは環境とインタラクションしつつタスク遂行に重要なセンサをセンサ値と報酬の相関から統計的に判断する。そして，重要なセンサを用いて Q 空間を再構築することで，従来よりも Q 空間を縮小することができる。これにより，学習データが削減され学習に要する時間が短縮する。以上のことを実現する手法を提案する。

これらロボットの外部と内部の情報の取捨選択によって，ロボットが利用する余分な情報を削減することができる。それにより効率的に学習が実現できることを示す。

ABSTRACT

At present, reinforcement learning is the most prominent learning method used when controlling an actual robot. A robot receives environmental information from its sensors as inputs and as outputs performs suitable actions. A robot needs to learn the relation between each input and output. A robot learns proper actions based on a learning space. The learning space consists of an input axis, an output axis, and an evaluation axis. When the number of sensors increases, the learning space expands and as a result, the time taken by a robot to learn a task increases.

The objective of this paper is to overcome this problem. If we reduce the learning space, the learning performance will also reduce. Therefore, I focus on reducing the learning time while keeping the learning space large. To achieve this, I follow two approaches. The first approach involves communicating with other robots, and gathering data for learning. Typically, a robot uses only the data it collects for learning. If the learning space is large, the time required by a robot to collect sufficient data increases. By using data collected from other robots, I attempt to accelerate the speed of learning. In Chapter 3, I examined an assumption with regard to the negative impact certain collected information could have on the robot. To this end, in Chapter 4, I propose a system in which a robot, when performing a task, selects only those robots that have profitable information. The second approach involves compressing the learning space by only considering sensors necessary to perform a task. Based on the task, some sensors are important and some are unimportant. By dynamic compression as per the task, I attempt to effectively accelerate the speed of learning. In

Chapters 5 and 6, I propose a method by which a robot statistically identifies important sensors through interaction with the environment.

In each chapter, I apply the proposed methods to the path planning problem. Two kinds of environment are used, maze and open space field. Experiments are performed using a computer simulation and an actual robot. In each case, I compare the proposed method with reinforcement learning and show the improvement of the learning speed with the high performance.

論文審査結果の要旨

本論文では,ロボットの行動学習アルゴリズムの高速化･高性能化を目的として，強化学習において環境情報を取捨選択する手法を新たに提案し，その有用性を実験により示している。

一般に，強化学習では Q 空間と呼ばれる学習空間に「環境情報」と「ロボット自身の行動」および「行動に対する評価」を「学習用データ」として蓄積して行動学習を行う。このうち環境情報はロボットが活動中にセンサ等を通じて環境から得る情報で，学習用データの量はセンサ数に依存する。ここで，学習能力の性能向上のためにセンサ数を増やし学習用データを増加させると，結果として Q 空間が拡大し学習時間が増大する。一方，センサ数を減らせば Q 空間が縮小し高速化が実現するものの，学習能力の性能が低下する。本論文では，この問題を解決し強化学習の高速化と高性能化の両立を図るために，二種類の「環境情報の取捨選択法」，すなわち「外部環境情報の取捨選択法」と「内部環境情報の取捨選択法」を提起している。

第一の「外部環境情報」の具体としては，着目するロボット（自ロボット）が他ロボットとのコミュニケーションにより獲得する他ロボットの経験情報を扱っている。他ロボットの経験情報を得られれば，自ロボットの活動から得られるデータより多くの学習用データを活用できる。しかし，自ロボットが直面する環境やタスクが他ロボットと異なる場合，他ロボットの経験情報は学習に不要なばかりか学習に悪影響を及ぼす。そのため本論文では，強化学習における報酬値に基づいて他ロボットの経験情報の重要度を算出し，自ロボットと類似した環境やタスクを持つ他ロボット空の経験情報のみを選出するアルゴリズムを提案している。

第二の「内部環境情報」の具体としては，自ロボットのセンサ情報を扱っている。搭載するセンサが少ないと環境認識能力が低下しタスク達成に支障をきたす。しかし単にセンサ数を増すと Q 空間が増大し学習時間の増大を招く。また，直面するタスクに不必要なセンサ情報はノイズとなり学習に悪影響を及ぼすこともある。そのため本論文では，強化学習における報酬値に基づいて直面する環境やタスクに応じた各センサ情報の重要度を算出し，その場面で必要となる自ロボットのセンサ情報のみを選出するアルゴリズムを提案している。

本論文ではさらに，提案した二手法を実装したロボットの行動学習システムを構築したうえでシミュレーション実験および実ロボットによる実験を行い，提案手法が実環境下で有用なことを実証的に示している。

本研究によって得られた知見は，ロボットの行動学習に関する研究分野に寄与するところが大であるとともに，本学及び専攻が定める審査基準を満たしており博士論文に値すると認められる。