Q-learning 算法及其在囚徒困境问题中的实现

张春阳　陈小平　刘贵全　蔡庆生
中国科学技术大学计算机系　(合肥 230027)
E-mail:zcywws@263.net

摘　要　Q-learning是一种优良的强化学习算法。该文首先阐述了Q-learning的基本学习机制,然后以囚徒困境问题为背景,分析、对比了Q-learning算法与TFT算法,验证了Q-learning算法的优良特性。

关键词　机器学习　强化学习　Q-learning算法　Agent　囚徒困境问题　针锋相对算法

文章编号 1002-8331-(2001)13-0121-02　**文献标识码** A　**中图分类号** TP301.6

Q-learning Algorithm and Its Usage in Prisoner's Dilemma

Zhang Chunyang　Chen Xiaoping　Liu Guiquan　Cai Qingsheng
(Department of Computer Science, University of Science & Technology of China, Hefei 230027)

Abstract: Q-learning is an algorithm of Reinforcement learning algorithm.In this paper,We elaborate the learning method of Q-learning algorithm simply,use it in prisoner's dilemma game and compare it with TFT to show its good character.

Keywords: Machine Learning, Reinforcement learning, Q-learning algorithm, Agent, prisoner's dilemma, TFT algorithm

1　强化学习(Reinforcement learning)与Q-learning算法

按照人工智能大师西蒙的观点,学习就是系统在不断重复的工作中对本身能力的强化或者改进,使得系统在下一次执行同样或类似任务时,会比现在的更好或效率更高。现在的计算机系统和人工智能系统的学习能力有限,因而不能满足科技和生产提出的新要求,所以机器学习成为人工智能的核心课题之一。[6]

机器学习按其信息来源可以分为有教师学习和无教师学习,按其所处的环境,可以分为静态环境中的学习和动态环境中的学习。在有教师学习中,需向学习器输入带标记信息,在静态环境中,这种带标记信息较易获得;而动态环境中的信息具有不完全性和不确定性,即某种行动将会导致什么结果并不能被确定,所以很难得到带标记信息。因此有教师学习不适于用在动态环境中。[4]

强化学习是一种无教师学习,其基本学习机制如图1所示。

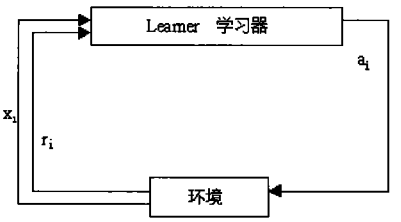

图1　强化学习的基本学习机制

此机制可以解释为:建立一个分时系统模型,在时间t=i时,输入向量x_i,这时选择行动a_i,得到一个收益r_i,$r_i=r(x_i,a_i)$。学习者的目标是找到一个策略$\Pi(X)$,使与输入向量对应的行动产生的收益和最大。

强化学习(RL)的基本学习机制表明,它能接受并处理动态环境中的不完全、不确定信息,产生最佳策略,选择最优行动,从而最大限度地影响其所处的动态环境。[1]

Q-learning算法是一种强化学习算法. 它的优势在于不需要对所处的动态环境建模,所以耗费时间少,能在Agent与环动态境交互时在线使用。

在Q-Learning算法中,Agent通过评价"状态---行为"对Q(s,a)来工作,Agent的经验包括一系列不同的状态或事件,其运行机制为:

·检测当前状态s;

·选择并执行行动a;

·测试后续状态s';

·获得一个立即收益r;

·利用学习因子α,调节Q的值

$$Q(s,a)=(1-\alpha)Q(s,a)+\alpha[r+\gamma\max_{b\in A}Q(s',b)]$$

初始Q(s,a)是随机给出的。

根据当前的Q值Q (s,a),Agent一般都选择能使Q(s,a)最大化的行动,然后用立即收益r和扣因子γ来调节Q值。如果用Q值来预测最终的收益,那么上述的学习过程就是TD(0)学习法。但是Q-learning强化了一般的TD(使用值迭代法运用在强化学习中)方法,这是因为TD要求前瞻一步,即要建立行动的影响模型,而Q-learning不需要。

一种用来表示Q值变化的简捷的符号为:

$$Q(s,a)\leftarrow\beta r+\gamma\max_{b\in A}Q(s',b)$$

基金项目:国家自然科学基金的支持(编号:69875017)

作者简介:张春阳,硕士研究生,主要研究领域为机器学习,Agent。陈小平,主任,博士,副教授,主要研究领域为机器学习,知识发现,逻辑系统。刘贵全,讲师,博士,主要研究领域为机器学习,知识发现。蔡庆生,教授,博导,主要研究领域为机器学习,知识发现。

这里 Q(s,a)是输入状态 s 和行动 a 对应的新 Q 值,r 是输入状态 s 采取行动 a 后得到的立即收益,s′为在状态 s 下,选择一步行动 a 所能到达的状态,β 是对新 Q 值的调节方法。

虽然对每个状态的最佳 Q 值的定义递归地依靠其后继 Q 值的计算,但是这些值并不需要提前计算。事实上,是通过随机机制产生后继状态,再用样本迭代法得到 Q 值。

Q-learning 是最好的通过计算 Q 值,产生概率,并根据概率随机地选择行动的学习方法。这是因为在一些特定的情况下,通过这种学习过程计算出的 Q 值会收敛于一些最佳的点(每个最佳的点都基于一个最佳的策略)。这是被 Watkins 和 Dayon 证明过的[1]。

2 Q-learning 与 TFT 算法的对比实验

2.1 囚徒困境问题与 TFT 算法

囚徒困境问题的陈述:两个犯罪同伙 A 和 B 被抓获,他们被单独审问。如果 A、B 互相出卖(Defect),他们将各得 1(P)分。如果 A、B 互相合作(Cooperate),他们将各得 3(R)分。如果 A 出卖 B,B 却未出卖 A,A 得 5(T)分,B 得 0(S)分。如果 B 出卖 A,A 却未出卖 B,B 得 5(T)分,A 得 0(S)分。

表 1 囚徒困境问题得分表

	Defect(A)	Cooperate(A)
Defect(B)	A=B=P	A=S,B=T
Cooperate(B)	A=T,B=S	A=B=R

囚徒困境问题是游戏理论发展中的一个早期游戏,它所具有的特点是一个使双方得分和最高的结果并不是使一方得分最高的结果,这就使囚徒困境问题成为能很好地研究社会中个人在冲突与合作之间怎样抉择的问题。

TFT 算法:即针锋相对算法,其基本策略就是以对手上一步的行动为当前行动[2]。

2.2 实验设计

如记载囚徒困境问题中当前状态前两步自己的走法,则一共有四个状态(CC,DC,CD,DD),每个状态又可以有两种走法 c 和 d,所以一共有 8 个 Q 值 Q(CC,c),Q(CC,d),Q(CD,c),Q(CD,d),Q(DC,c),Q(DC,d),Q(DD,c),Q(DD,d)需要记载,图 2 为它们之间的状态转换图。

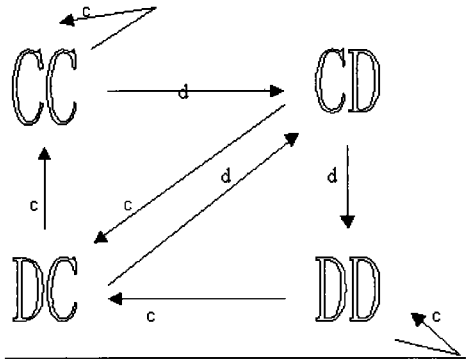

图 2 囚徒困境问题状态转换图

1.初始化 Q 值为 0;

2.随机产生前两步 a_1 和 a_2;

3.利用 a_{n-2} 和 a_{n-1} 确定状态 s,再利用公式

$$P(a)=e^{Q(s,a)/t}\Big/\sum_{a\in A}e^{Q(s,a)/t}$$

确定在状态 s 下走 c 的概率和走 d 的概率,然后根据概率随机产生 a_n 的走法。

4.用公式

$$Q(s,a)=(1-\alpha)Q(s,a)+\alpha[r+\gamma\max_{b\in A}Q(s',b)]$$

改变 Q(s,a)的值。

5.回到 3。

2.3 结果分析

说明:对不同的 γ 统计结果,每个 γ 有 12 组,每组为 998 次对抗,记录总得分情况和 Q-learnging 选择合作的次数:

表 2 Q-learning 算法 VS TFT 算法结果统计

γ	Q-learning 算法得分	TFT 算法得分	Q-learning 算法选择 C 的次数
0.1	24786	24746	4814
0.3	26321	26281	5523
0.5	27230	27200	5972
0.7	28136	28106	6450
0.9	29926	29901	7340

表 3 Q-learning 算法 VS 随机算法结果统计

γ	Q-learning 算法得分	随机算法得分	Q-learning 算法选择 C 的次数
0.1	26949	21840	4609
0.3	28930	21948	4546
0.5	29096	21581	4403
0.7	29191	21336	4372
0.9	32640	13490	2113

1.在 Q-learning 算法 VS 针锋相对算法的结果统计中可发现以下规律。

A.Q-learning 算法的得分与 TFT 算法的得分十分接近,但 Q-learning 算法的得分始终不低于 TFT 算法的得分,这说明 Q-learning 算法是一种行之有效的对付 TFT 算法的算法,而 TFT 算法一直被认为是解决囚徒困境问题的最佳算法,这说明了 Q-learning 算法的优越性。

B. 从统计结果上可看出,随着折扣因子 γ 值的增加,Q-learning 算法选择合作(Cooperate)的概率越大,说明在对付 TFT 算法时,过去状态对当前状态影响越大,则状态 Q(S,c)的值相应增长快,Q-learning 算法选择合作的概率就高。

C. 当 Q-learning 算法选择合作的概率增大时,Q-learning 算法的得分与 TFT 算法的得分都增加,这说明合作是这两个算法在对抗中的较优选择。

2.在 Q-learning 算法 VS 随机算法(行动随机产生)的统计结果中可发现以下规律。

A.Q-learning 算法的平均得分远远高于随机算法的得分,再次说明了 Q-learning 算法在囚徒困境问题中的优势。

B.与 Q-learning 算法 VS 针锋相对算法相反,随着折扣因子 γ 值的增大,在 Q-learning 算法 VS 随机算法中,Q-learning 算法选择合作的概率却减小了,这说明在对付不同的算法时 Q-learning 算法可以选择不同的策略,即 Q-learning 有良好的适应性。

C. 随着 Q-learning 算法选择合作的概率减少,Q-learning 算法得分与随机算法得分的差距加大,说明在 Q-learning 算法 VS 随机算法中,出卖对方(Defect)是较优的选择。

在 Q-learning 算法 VS 针锋相对算法和 Q-learningVS 随机算法中有一个共同的特点,即折扣因子越大时 Q-learning 算法所的的分数越高,说明了在 Q-learning 中当前状态对过去状态的依赖性较高。

(下转 128 页)

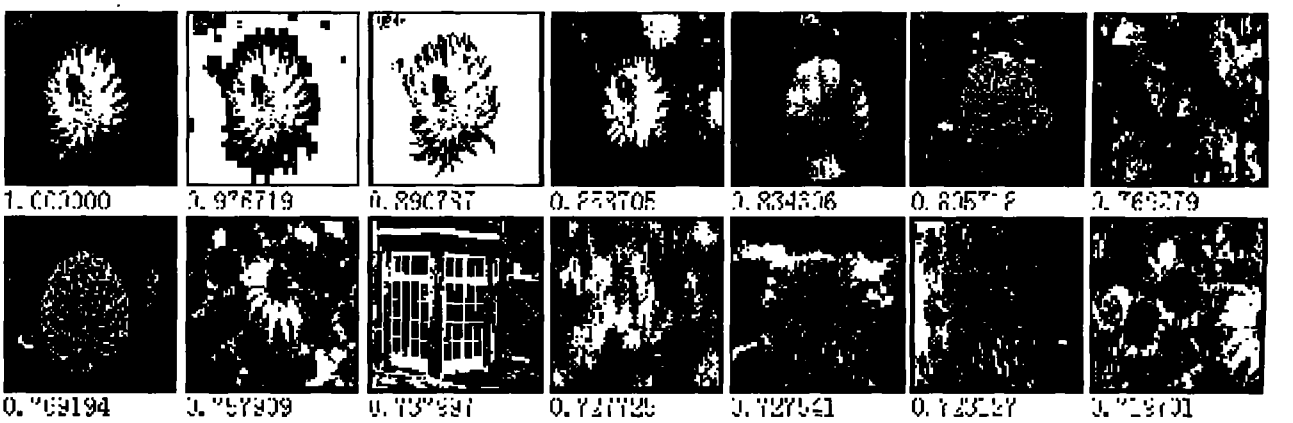

图 1　该算法检索结果 t_α=0.6，t_β=0.4

结果，则检索精度得到了很大的提高。

设相交颜色直方图相交算法的相似度比例因子为 t_α，考虑两图象中相交颜色各自按比例分布情况的相似度比例因子为 t_β，则综合两种算法得到图象 Q、I 的相似度为：

$$Sim(Q,I)=t_\alpha \times Sim_\alpha + t_\beta \times Sim_\beta$$

4　实验结果

系统采用 Visual C++ 6.0 编程语言在 Windows 98 环境下实现，对一个包含近 500 幅彩色图象的数据库做实验，这些图象全部来自网上的 QBIC 图象数据库。下图为算法的最后检索结果，其中第一幅图为示例图象，第二、三幅图为第一幅图改变了背景颜色之后的图象。该文算法具有计算复杂度低、检索速度快的优点，从检索结果看，算法获得了较好的检索精度，并可用于检索背景颜色不同的图象。（收稿日期：2000 年 6 月）

参考文献

1.Finn R.Query by image content[J].IBM Research,1996;3:22~25

2.Flickner M et al.Query by Image Video Content:The QBIC System [J].Computer,1995;23~32

3.J R Smith,S-F Chang.VisualSEEK:a fully automated content-based image query system[R].Technical report,Columbia University,1996.9

4.Jain A K,Vailaya A.Image retrieval using color and shape[J].Pattern Recognition,1997;29(8):1233-1244

5.Swain M,Ballard D.Color indexing[J].International Journal of Computer Vision,1991;7(1):11-32

6.Mehtre B M,Kankanhalli M S,Narsimhalu A D et al.Color matching for image retrieval[J].Pattern Recognition Letters,1995;16(3):325-331

7.Healey G,Slater D.Global Color Constancy:Recognition of Objects by Use of Illumination Invariant Properties of Color Distributions[J].J Opt Soc Am A,1994;11(11):3003~1010

8.Seaborn,M.On the use of colour in content based image retrieval[C].Proc.ACM Conf.1997

9.John R Sith,Shih-Fu Chang.Automated Image Retrieval Using Color and Texture[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1996.11

10.John R Smith,Shih-Fu Chang.Tools and Techniques for Color Image Retrieval[J].SPIE 1996;2670:426~437

11.John R Smith,Shih-Fu Chang.Local color and texture extraction and spatial query[C].Proc.of IEEE Int Conf.Image Processing,1996

12.John R Smith.Integrated Spatial and Feature Image Systems:Retrieval,Analysis and Compression[D].Thesis of the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences,Columbia University,1997

13.J R Smith,S-F Chang.Single color extraction and image query[R].Technical report,Columbia University,1995

（上接 122 页）

3　结论

强化学习是机器学习的一种重要方法，特别是在它的设计者不能预见此 Agent 所能遇到的所有情况的时候，而这种状况却是 Agent 经常遇到的。Q-learning 算法通过计算得出下一步选择每一种行为的概率，并根据概率随机产生下一步行动。这种强化学习在以囚徒困境问题为背景的上述实验中，优于以前所认为解决此问题的最好算法 TFT 算法，这是因为在一些特定的情况下，通过 Q-learning 学习过程计算出的 Q 值会收敛于一些最佳的点，而每个点对应于一个最佳策略。但在一些情况下，如在囚徒困境问题中对手一直选择合作，则 Q-learning 的策略会失败，即在这种情况下并没有收敛于最佳的点，怎样解决这样的问题，将是作者下一步的工作。

(收稿日期：2000 年 5 月)

参考文献

1.Leslie Pack Kaelbling,Andrew W Moore.Reinforcement learning:A survey[J].Journal of AI Research,1996

2.Tuomas W Sandholm,Robert H Crites.Multiagnet Reinforcement Learning in the Iterated Prisoner's Dilemma.

3.Sandip Sen,Mathendra Sekaran.Individual learning of coordination knowledge.1996

4.John Wilbar Sheppard.Multi-Agent Reinforcement Learning in Markov Games.1997

5.P R 科恩,E A 费根鲍姆 编著,周少柏,董汛 译.人工智能手册[M].第三卷,科学出版社,1991

6.蔡自兴,徐光佑.人工智能及其应用[M].第二版,清华大学出版社,1996

7.史忠植.高级人工智能[M].科学出版社,1998

8.P Cichosz.Reinforcement Learning Algorithms Based on the Method of Temporal Differences.1994