doi:10.3969/j.issn.1673-4785.2009.03.003

回报函数学习的学徒学习综述

金卓军,钱 徽,陈沈轶,朱淼良
(浙江大学 计算机学院,浙江 杭州 310027)

摘 要:通过研究基于回报函数学习的学徒学习的发展历史和目前的主要工作,概述了基于回报函数学习的学徒学习方法.分别在回报函数为线性和非线性条件下讨论,并且在线性条件下比较了2类方法——基于逆向增强学习(IRL)和最大化边际规划(MMP)的学徒学习.前者有较为快速的近似算法,但对于演示的最优性作了较强的假设;后者形式上更易于扩展,但计算量大.最后,提出了该领域现在还存在的问题和未来的研究方向,如把学徒学习应用于POMDP环境下,用PBVI等近似算法或者通过PCA等降维方法对数据进行学习特征的提取,从而减少高维度带来的大计算量问题.

关键词:学徒学习;回报函数;逆向增强学习;最大化边际规划
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-4785(2009)03-0208-05

Survey of apprenticeship learning based on reward function learning

JIN Zhuo-jun, QIAN Hui, CHEN Shen-yi, ZHU Miao-liang
(Department of Computer Science, Zhejiang University, Hangzhou 310027, China)

Abstract: This paper focuses on apprenticeship learning, based on reward function learning. Both the historical basis of this field and a broad selection of current work were investigated. In this paper, two kinds of algorithm—apprenticeship learning methods based on inverse reinforcement learning (IRL) and maximum margin planning (MMP) frameworks were discussed under respective assumptions of linear and nonlinear reward functions. Comparison was made under the linear assumption conditions. The former can be implemented with an efficient approximate method but has made a strong supposition of optimal demonstration. The latter has a relatively easy to extend form but may take large amounts of computation. Finally, some suggestions were given for further research in reward function learning in a partially observable Markov decision process (POMDP) environment and in continuous/high dimensional space, using either an approximate algorithm such as point-based value iteration (PBVI) or a feature abstraction algorithm using dimension reduction methods such as principle component analysis (PCA). Adopting these may alleviate the curse of dimensionality.

Keywords: apprenticeship learning; reward function; inverse reinforcement learning; maximum margin planning

学徒学习(apprenticeship learning),又称为示教学习(imitation learning)、模仿学习(imitation learning)或者观察学习(learning by watching)等.它是指学习者模仿专家的行为或者控制策略的过程[1].文献[2]介绍了基于边际最大化的学徒学习的发展过程,其重点放在MMP(maximum margin planning)框架的提出、解决及优化.本文在文献[3]的基础上,对基于逆向增强学习的学徒学习和MMP框架2类方法的作了更具体的分析和比较,另外讨论了近几年来该领域的一些最新进展.

在移动机器人控制当中,规划模块将回报函数作为输入,然后算出一条使回报函数值最大的决策序列,这种方法已经成为自动移动机器人系统的核心.但是,要建立回报函数的过程在实际操作中是比较困难的,回报函数常常需要手工调节,然后观察运行结果,继而再调整回报函数,如此迭代来完成回报函数的构建.这样的过程在现实情况中往往是不可取的.

基于回报函数学习的学徒学习被用来从专家演

收稿日期:2008-10-08.
基金项目:国家自然科学基金资助项目(90820306);浙江省科技厅重大资助项目(006c13096).
通信作者:钱 徽. E-mail: qianhui@zju.edu.cn.

示的策略序列中近似还原出恰当的回报函数,之后便可以用传统的规划方法寻求最优策略. 目前用于从示教中还原回报函数的主要方法有基于逆向增强学习的学徒学习和 MMP 框架 2 种. Ng 和 Russell 提出了逆向增强学习(inverse reinforcement learning, IRL)[4],它通过最大化专家演示策略和其他策略的差别,还原出一个能得出和专家演示相似策略的回报函数. Abbeel 等人将逆向增强学习进行拓展,称为学徒学习[5],并且在文中以一个驾驶模拟实验系统证明了该算法可以快速通过学徒学习掌握不同的驾驶风格. 近年来,Kolter 等人又将层次的概念引入学徒学习[6],并首次应用于四足机器人的实验中. 另一方面,Ratliff 等人通过将该问题转化为二次最优化问题,并由之提出了 MMP 框架[2],同时在此基础上设计了一系列基于梯度下降的算法[7]. 近年来,Syed 等人又从线性规划[8]和博弈论角度[9]讨论了该问题,将该问题转化成线性规划问题,运算效率得到大大提高;但在专家示教是否为最优未知时,作者没有提出一个统一的有效算法. 文献[9]中将学习过程解释为 2 个玩家间的零和博弈,此方法改进了基于逆向增强学习的学徒学习在专家示教非最优情况下学习结果不理想的缺点. 另外,为了将该算法应用到实际情况中,Grimes 和 Rao 等人在文献中[10]中探讨了在不确定环境下的学徒学习系统设计.

目前为止,基于回报函数学习的学徒学习已经被应用到如直升机特技表演[11],四足机器人的复杂地形穿越及机器手臂控制等领域[6, 12-13],并且取得了良好的效果.

1　模型及符号约定

本文介绍的方法都是基于马尔可夫决策过程(Markov desision process, MDP)模型的,关于 MDP 模型及基于 MDP 的机器学习可以参考相关文献[14-15],在此仅以符号约定为目的简述如下:

MDP 模型可以表示为五元组

$$(S, A, P_{sa}(\cdot), \gamma, R).$$

式中:S 是有限个状态的集合,设状态个数为 N;$A = \{a_1, \cdots, a_k\}$ 是 k 个动作的集合;$P_{sa}(\cdot)$ 是在状态 s 中执行动作 a 后的转移概率;γ 是折合因子,范围在 $[0,1]$ 之间;R 是状态到实数集的映射,表示状态所对应的回报值.

策略 $\pi \in R \mapsto A$,某策略的值函数可表示为

$$V^{\pi}(s_1) = \mathrm{E}[R(s_1) + \gamma R(s_2) + \gamma R(s_3) + \cdots \mid \pi].$$

Q 函数定义为

$$Q^{\pi}(s,a) = R(s) + \gamma E_{s' \sim P_{sa}(\cdot)}[V^{\pi}(s')].$$

Bellman 方程可表示为

$$V^{\pi}(s) = R(s) + \gamma \sum_{s'} P_{s\pi(s)}(s') V^{\pi}(s'),$$

$$Q^{\pi}(s) = R(s) + \gamma \sum_{s'} P_{sa}(s') V^{\pi}(s').$$

学习特征是状态的组合,是代表某一特征的状态的集合. 学习特征的基回报函数简称为基回报函数,是使符合某一学习特征的策略回报值最高的回报函数. 基回报函数形式化表达为 $f_i \in S \times A \mapsto R^d$,$d$ 为学习特征的个数. 回报函数 $R: S \times A \mapsto R$,代价函数是负的回报函数 $c: S \times A \mapsto R$.

2　基于线性回报函数的学习

在基于线性回报函数的假设下,线性回报函数是基于回报函数的线性组合:

$$R(s) = w_1 f_1(s) + w_2 f_2(s) + \cdots + w_d f_d(s).$$

式中:$f_1, \cdots, f_d$ 是确定的基回报函数(base function),每一个学习特征对应一个基回报函数;$\boldsymbol{w} = [w_1\ w_2 \cdots w_d]$ 为各个基回报函数之间的权值向量.

下面介绍 2 种基于以上假设的学徒学习的方法,它们分别是基于逆向增强学习的学徒学习和基于 MMP 框架的学徒学习. 针对回报函数为非线性假设的算法将在后面提及.

2.1　逆向增强学习

在逆向增强学习中[4],算法通过专家的演示得到回报函数,它假设专家是基于一个能产生最优或者近似最优策略的回报函数来进行演示的. 学习者没有必要也不可能找出真实的回报函数,因为要找出真实回报函数是一个数学病态问题,因此,学习者只需近似"还原"出适当的回报函数.

Ng 和 Russell 提出逆向增强学习后[4],Abbeel 和 Ng 将增强学习引入学徒学习[5],策略 π 对应的值函数可以表示成:

$$V_{\boldsymbol{w}}(\pi) = \boldsymbol{w}^{\mathrm{T}} \cdot \mathrm{E}[\sum_{t=0}^{\infty} \gamma^t \varphi(s_t) \mid \pi].$$

式中:γ 为折合因子;等式右边除了 $\boldsymbol{w}$ 以外的部分称为特征期望,记为 $\boldsymbol{\mu}$,作为算法中 2 种策略之间相近程度的衡量标准.

逆向增强学习通过使执行专家演示策略和次优策略时获得的回报值的差最大来求得各特征之间的权值 $\boldsymbol{w}$,因此,该学习问题可以归结为以下的最优化问题:

$$\max_{\tau, \boldsymbol{w}: \|\boldsymbol{w}\|_2 \leqslant 1} \tau,$$

$$\text{s.t. } V_{\boldsymbol{w}}(\pi_E) \geqslant V_{\boldsymbol{w}}(\pi_i) + \tau, i = 1, \cdots, t-1. \quad (1)$$

式中:π_E 为专家演示策略. π_i 为已有的第 i 次迭代产生的策略.

文献[5]中提出了基于逆向增强学习的学徒学习迭代算法,并且提出 2 种方法,分别为边际最大化方法(max-margin method)和投影法(projection method). 前者将专家策略的回报期望与目前次优策略回报期望的

距离作为目标函数,然后可以通过 SVM 方法来求解;而后者是一种更为简单的近似计算方法,其优点是避免了求解二次最优化问题. 由于专家示教的不准确性,专家演示的最优条件较难达到, Coates 等人提出通过多次次优演示来提取最优策略的方法[16].

2.2 MMP 框架

Taskar 首先提出了结构最大边际(structured large margin)框架[17], 在此框架下,Ratliff 等人将学徒学习归结为在策略空间上的边际最大化结构化预测问题(structured prediction problem)[18]. 在这种方法当中,学习的对象是从特征回报函数到代价函数的映射. 学习者的目标是接收示例策略,然后通过学习做出相同或者类似的决策. 为了解决这个问题, Ratliff 等人提出了 MMP 框架.

由于 Ratliff 等人在提出该方法时学习的是代价函数,因此这里沿用了这一概念;但在实际中,代价函数和回报函数的意义是类似的,因为代价函数可看作是负的回报函数. 为了还原代价函数,需要不断调整代价函数来使得示例路径看起来是最优的. 算法的目标就是寻找一个这样的代价函数,在使用这个代价函数时,专家给出的示例路径有最小的代价值. 因此,该算法把示例路径的代价值和拥有最小代价的路径的代价之差作为优化程序的衡量标准. 上面的问题可以用机器学习中的 MMSC(maximum margin structured classification) 来描述.

当示例的策略和其他的策略相差较小时,算法需要调整 $\boldsymbol{w}$ 来使之增大;但是为了使 $\boldsymbol{w}$ 的值有意义,该优化问题又要优先考虑值较小的 $\boldsymbol{w}$. 综合以上考虑,该问题可写成凸函数最优化问题:

$$\min_{\boldsymbol{w}\in\boldsymbol{W}} R(\boldsymbol{w}),$$

$$R(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{F}_i\boldsymbol{\mu}_i - \min_{\boldsymbol{\mu}\in\boldsymbol{G}_i}\{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{F}_i\boldsymbol{\mu} - \boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{\mu}\}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2. \tag{2}$$

式中:$\boldsymbol{u}\in R_+^{|R||A|}$ 代表每个状态-动作对所对应的访问频率;$\boldsymbol{F}\in R^{d\times|S||A|}$ 为基回报函数的矩阵;$\boldsymbol{G}_i$ 为 $\boldsymbol{\mu}$ 所在的空间;$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{F}_i\boldsymbol{\mu}_i - \min_{\boldsymbol{\mu}\in\boldsymbol{G}_i}\{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{F}_i\boldsymbol{\mu} - \boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{\mu}\} = \xi_i$ 为引入的一组松弛变量,代表每个示例中违反约束时的惩罚;$\boldsymbol{l}_i\in R_+^{|S||A|}$ 表示在示例 i 中的损失函数;$\lambda\geqslant 0$ 是一个用于正规化的参数. 式(2)的优化目标是在违反最少的约束条件下找到尽可能小的 $\boldsymbol{w}$.

MMP 模型的数值计算方法采用了子梯度方法[19]的改进方法[7],其他的计算方法还包括 cutting plane[20]和 extra-gradient[18]方法等. Ratliff 等人还结合 boosting 思想提出了 MMPBOOST 算法,使原算法能够很好地适应新的学习特征. 另外,文献[21]研究了在某些学习特征缺失的情况下的学徒学习.

2.3 基于 IRL 和 MMP 的学徒学习的比较

基于 IRL 的学徒学习是一种用不同基回报函数的线性组合来逼进真实回报函数的迭代过程. 而 MMP 方法可以看作是二次最优化问题的基于梯度下降求解法[19]. 2 类方法的处理过程如图 1 和图 2 所示.

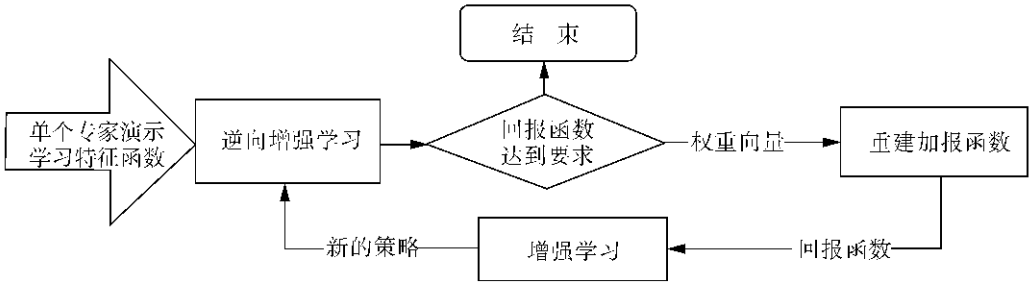


图 1 基于逆向增强学习的学徒学习

Fig. 1 Apprenticeship learning based on IRL

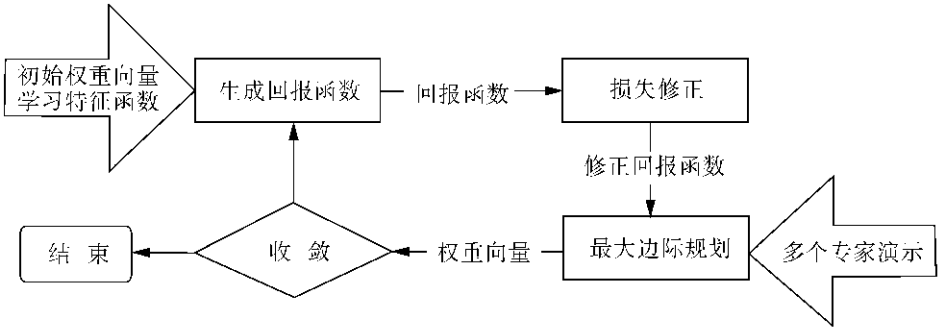


图 2 基于 MMP 的学徒学习

Fig. 2 Apprenticeship learning based on MMP

这2种方法的共同处在于2点:一是它们都将学习问题转化为对二次最优化问题的求解,优化模型的基本思想都是使最优策略与其他非最优策略的边际最大化,即使得专家演示和其他策略所获得的回报之差尽可能大.二是它们的学习的目标相同,即通过学习各学习特征之间的权重向量来还原回报函数.

这2种方法的区别类似于生成式学习和判别式学习[2].首先,基于IRL的学徒学习只针对一个MDP模型,MMP方法的专家示例则可以来自多个不同的MDP模型,这些MDP模型可以有各自的状态、动作及转移矩阵,它们之间通过学习特征向量(feature vectors)来取得统一.基于IRL的学徒学习假定专家的演示是最优或者近似最优的,或者说它假定存在一个使得专家产生最优策略的回报函数,而专家则是根据它来进行演示的.这样的假设条件相对较强,而MMP的假设则弱得多.其次,二者对每一轮迭代中回报函数的更新方法不同.基于IRL的方法是通过将新产生的策略加入到已有策略集合,然后通过最优化方法求出下一轮的回报.MMP过程中采用根据专家演示修正回报函数的方法,将专家演示所涉及到的学习特征对应的回报增加,其余的减小,且保证离专家演示越远的策略回报减小越多,从而更新回报函数.最后,二者解最优化模型采用的方法不同,前者用的是SVM和投影法,后者用的是梯度下降法.

基于逆向增强学习的学徒学习算法目前还有以下缺点:1)它对于基回报函数的设计非常敏感.这使得该算法过于依赖人为的基回报函数的设计.尽管PCA等方法可以被用来从状态中提取学习特征,但在实际应用中这会使算法在特征的提取上花很多时间.2)该算法的不足之处在于缺乏那些部分专家演示较少访问到的状态相关信息,从而导致学习结果在某些状态不理想.文献[22]针对这2个问题进行了研究,并结合梯度算法给出了解决方案.3)该算法的局限性来自于上面提到的线性假设.而下面的基于MMP的改进算法则去掉了代价函数的线性假设.

MMP框架可以被扩展到非线性回报函数的情况,Ratliff等人基于boosting的泛函梯度下降理论建立了指数梯度下降的一般化算法,称为LEARCH算法.

LEARCH用更一般化的形式$\boldsymbol{c}(f_i^{sa})$来表示代价函数,这样式(2)中的$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{F}_i\boldsymbol{\mu}$就可以写为$\sum_{(s,a)\in M_i}\boldsymbol{c}(f_i^{sa})\boldsymbol{\mu}^{sa}$.式(2)就可以写成下面的形式:

$$R(\boldsymbol{c})=\frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{(s,a)\in M_i}\boldsymbol{c}(f_i^{sa})\mu_i^{sa}-\min_{\boldsymbol{\mu}\in G_i}\Big\{\sum_{(s,a)\in M_i}(\boldsymbol{c}(f_i^{sa})\boldsymbol{\mu}^{sa}-\boldsymbol{l}^{sa})\boldsymbol{\mu}^{sa}\Big\}\Big).\quad(3)$$

3　结束语

学徒学习让人们摆脱通过反复实验手动调节代价函数的烦琐过程,使学习者可以通过学习专家的演示来学习最优代价函数,从而生成最优策略.本文的主要探讨对象为基于回报函数学习的学徒学习方法,这也是目前学徒学习的主要方法.文中在回报函数为基回报函数的线性组合和回报函数为非线性2种假设下分别作了概述,主要介绍了基于IRL和MMP框架的学徒学习,并且比较了它们的优缺点.

基于回报函数学习的学徒学习中存在许多亟待解决的问题.首先应考虑非完全观察状态下的学徒学习.上面的方法都是建立在MDP模型上的,然而在实际应用中情况往往是非完全观察状态下的,即是建立在POMDP上的学徒学习.其次,在上面的方法中,学习特征的提取和设计都是人为完成的,这样就导致结果受学习特征的选择影响很大,事实上学习特征有可能通过对状态空间用PCA等方法降维来得到.另外,高维空间下学徒学习所带来的巨大计算量成为该学习方法被应用到更广泛的领域的主要障碍,因此,设计高维状态空间下更有效的求解方法也成为人们关心的问题之一.

参考文献:

[1]ATKESON C G, SCHAAL S. Robot learning from demonstration[C]//Proceedings of the Fourteenth International Conference on Machine Learning. Nashville, USA, 1997: 12-20.

[2]RATLIFF N D, BAGNELL J A, ZINKEVICH M A. Maximum margin planning[C]//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 729-736.

[3]金卓军, 钱　徽, 陈沈轶,等. 基于回报函数逼近的学徒学习综述[J]. 华中科技大学学报:自然科学版,2008(S1): 288-290, 294.
JIN Zhuojun, QIAN Hui, CHEN Shenyi, et al. Survey of apprenticeship learning based on reward function approximating[J]. Journal of Huazhong University of Science and Technology: Nature Science, 2008, 36(S1): 288-290, 294.

[4]NG A Y, RUSSELL S J. Algorithms for inverse reinforcement learning[C]//Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco, USA, 2000: 663-670.

[5]ABBEEL P, NG A Y. Apprenticeship learning via inverse reinforcement learning[C]//Proceedings of the Twenty-first International Conference on Machine Learning. Banff, Canada, 2004:1-8.

[6]KOLTER J Z, ABBEEL P, NG A Y. Hierarchical apprenticeship learning with application to quadruped locomotion [C]//Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2008.

[7]RATLIFF N, BAGNELL J A, ZINKEVICH M A. Subgradient methods for maximum margin structured learning[C]// Workshop on Learning in Structured Outputs Spaces at ICML. Pittsburgh, USA, 2006.

[8]SYED U, BOWLING M, SCHAPIRE R E. Apprenticeship learning using linear programming[C]//Proceedings of the 25 International Conference on Machine Learning (ICML 2008). Helsinki, Finland, 2008: 1032-1039.

[9]SYED U, SCHAPIRE R E. A game-theoretic approach to apprenticeship learning[C]//Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2008.

[10]GRIMES D B, RAJESH D R, RAO R P N. Learning nonparametric models for probabilistic imitation [C]//Proceedings of Neural Information Processing Systems. Cambridge, USA: MIT Press, 2007: 521-528.

[11]ABBEEL P, COATES A, QUIGLEY M, et al. An application of reinforcement learning to aerobatic helicopter flight[C]//Proceedings of Neural Information Processing Systems. Cambridge, USA: MIT Press, 2007: 1-8.

[12]KOLTER J Z, RODGERS M P, NG A Y. A complete control architecture for quadruped locomotion over rough terrain[C]//IEEE International Conference on Robotics and Automation. Pasadena, USA, 2008: 811-818.

[13]REBULA J R, NEUHAUS P D, BONNLANDER B V, et al. A controller for the littledog quadruped walking on rough terrain[C]//2007 IEEE International Conference on Robotics and Automation. Roma, Italy, 2007: 1467-1473.

[14]KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285.

[15]SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge, USA: MIT Press, 1998.

[16]COATES A, ABBEEL P, NG A Y. Reinforcement learning with multiple demonstrations [C]//The Twenty-first Annual Conference on Neural Information Processing Systems (NIPS 2007). Vancouver, Canada, 2007.

[17]TASKAR B, CHATALBASHEV V, KOLLER D, et al. Learning structured prediction models: a large margin approach[C]//Proceedings of the 22nd International Conference on Machine Learning. New York, USA: ACM, 2005: 896-903.

[18]TASKAR B, LACOSTE-JULIEN S, JORDAN M. Structured prediction via the extragradient method [C]//Proceedings of Neural Information Processing Systems. Vancouver, Canada, 2005: 1345-1352.

[19]SHOR N Z, KIWIEL K C, RUSZCAYNSKI A. Minimization methods for non-differentiable functions [M]. New York, USA: Springer-Verlag, 1985.

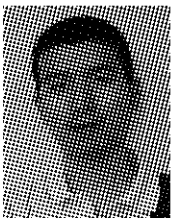
[20]TSOCHANTARIDIS I, JOACHIMS T, HOFMANN T, et al. Large margin methods for structured and interdependent output variables[J]. The Journal of Machine Learning Research, 2005, 6: 1453-1484.

[21]CHECHIK G, HEITZ G, ELIDAN G, et al. Max-margin classification of incomplete data [C]//Advances in Neural Information Processing Systems: Proceedings of the 2006 Conference. Cambridge, USA: MIT Press, 2007: 233-240.

[22]NEU G, SZEPESVARI C. Apprenticeship learning using inverse reinforcement learning and gradient methods[C]// Proceedings of Uncertainty in Artificial Intelligence. Vancouver, Canada, 2007: 295-302.

作者简介:

金卓军,男,1984 年生,博士研究生,主要研究方向为机器学习.



钱 徽,男,1974 年生,副教授,人工智能学会智能机器人专业委员会委员,主要研究方向为人工智能、计算机视觉.

陈沈轶,男,1980 生,博士研究生,主要研究方向为机器学习.