**Abstract**

温度参数在基于Softmax函数的动作选择中起重要作用用于将原始矢量变换为概率矢量。一种名为Opti Softmax的有效方法，用于确定加固学习中Softmax函数的最佳温度参数是本文开发的。 首先，设计了一种新的评估函数来衡量其有效性温度参数考虑信息损失的转换和多样性概率向量元素。其次,导出迭代更新规则以确定最优温度参数通过计算评估函数的最小值。最后，实验合成数据和D武装强盗问题的结果证明了其可行性和有效性Opti-Softmax方法。

**Introduction**

Softmax函数是归一化指数函数[4]将具有任意实数值的D维原始矢量转换为具有实数值的D维概率向量范围[0,1]加起来1.通常应用Softmax函数到机器学习领域，如逻辑回归[5]，

人工神经网络[15]，强化学习[17]。 一般而言，使用没有温度参数的Softmax功能在逻辑回归和多元分类问题中人工神经网络的最后一层，而具有温度参数[17]的Softmax函数用于转换动作奖励强化学习中的动作概率。

温度参数是Softmax行为中勘探 - 开发权衡的重要学习参数选择。大温度参数将导致

仅探测状态（动作具有几乎相同的概率选择），而小温度参数将导致仅利用状态（具有更高的行为）奖励更容易被选中）。本文重点介绍Softmax基于功能的探索 - 开发权衡在D武装强盗[21]的情景中，这是一个经典的行动选择问题强化学习。一些有代表性的研究涉及

Softmax动作选择总结如下。 [12]中的Koulourio-tis和Xanthopoulos研究了Softmax算法

温度参数0.3。 [19]中的Tokic和Palm测试了具有温度参数0.04,0.1,1,10,25和100的Softmax动作选择算法的性能。在[13]中，Kuleshov和Precup进行了彻底的实证比较最受欢迎的多臂强盗算法，包括Softmax温度参数为0.001,0.007,0.01,0.05和0.1。关于Softmax动作选择的其他研究可以是见文献[1,6,8,11,16,18]。据我们所知，现有研究主要采用试错策略进行选择

用于处理时Softmax功能的温度参数D武装强盗问题。

简单的Softmax功能将是一个非常有效的行动选择策略，以解决D型武装强盗问题，如果适当的话温度参数可以自动确定。取决于现在，没有提供这种自动温度的研究处理D型武装匪徒问题时Softmax功能的参数选择。因此，我们开发了一个有用的方法Opti-Softmax确定最佳温度参数Softmax功能在本文中。首先，我们设计了一个新的评估测量温度参数的有效性的功能。评估功能包括两部分：信息丢失原始向量与概率向量之间的关系概率向量元素之间的多样性其次，我们推导出迭代更新规则以确定最佳温度参数通过计算评估函数的最小值。

最后，对合成数据和D武装强盗问题进行了必要的实验，以验证其性能我们提出Opti-Softmax方法。本文的其余部分安排如下。第2节陈述了Softmax功能的问题表述。第3节给出了Opti-Softmax方法来确定最佳温度参数用于Softmax功能。一些实验模拟给出了最后，第5节给出了本文的简要结论。

**Problem formulations of Softmax function**

**(Softmax功能的问题公式)**

给定D维原始矢量x =（x1，x2，...，xD），xd∈R，d = 1,2，....，D并且存在k∈{1,2，。。 。，D}使得xk0，它可以转换成具有以下Softmax函数的D维概率向量p =（p1，p2，...，pD）：

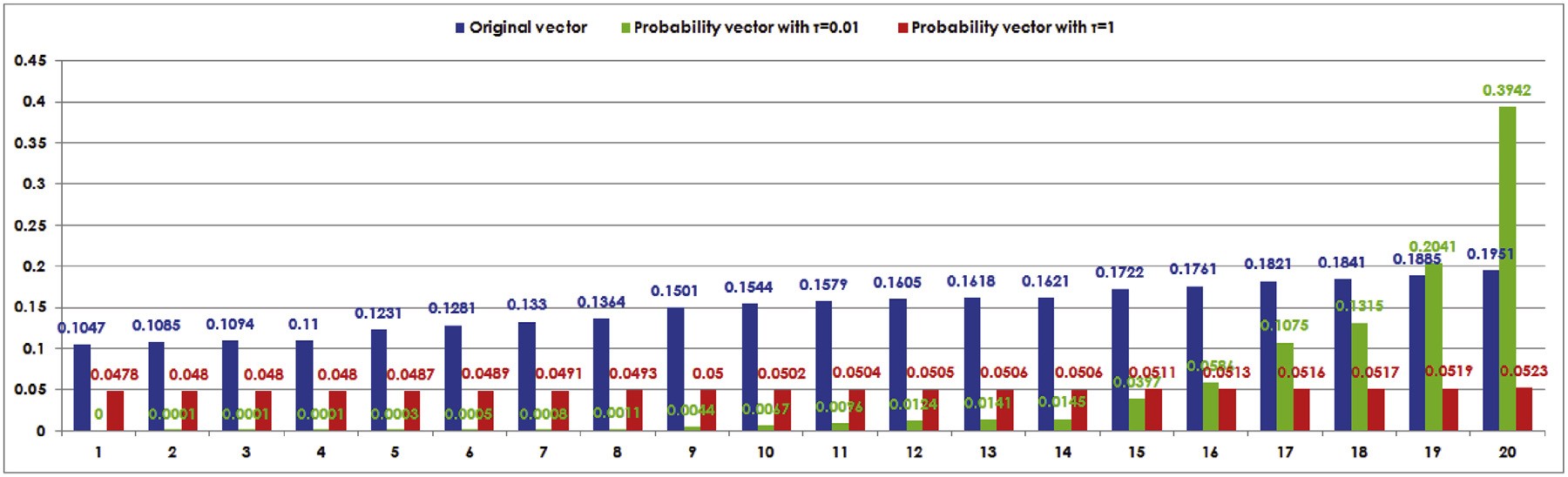


其中pd∈（0,1），和> 0是温度参数，它对转换有重要影响Softmax功能的性。

当→+∞, pd → ，pds的多样性很小；当→0  即，pds之间的差异很大。 图1提供了一个示例来显示Softmax函数的影响。属于区间[0.1,0.2]的20个实数（蓝条）是随机生成。 在这个图中，我们可以看到对应于= 1（红色条）的概率向量元素几乎是= 0.05，而= 0.01制作一些概率向量元素（绿色条）接近0。

在强化学习中，可以使用Softmax函数选择D-armed bandit中的bandit-arm，每个bandit-arm

为赌徒提供随机奖励。 假设=（x1，x2，...，xD）是与D-armed bandit对应的奖励向量。 然后，pd是选择第d个bandit-arm的概率。 →+∞ 将会导致（几乎选择相同的概率的bandit-arms），而→0将会导致（具有较高奖励的bandit-arm更容易选择）。 解决D-armed bandit的关键是如何选择bandit-arm以便赌徒能够获得最大的奖励。

 （**图一**）

**3. Determination of the optimal temperature parameter for Softmax function**

**(确定Softmax函数的最佳温度参数)**

本节介绍一种名为Opti-Softmax的新方法确定Softmax的最佳温度参数Opti功能。 首先设计以下评估函数L（）测量温度参数的有效性：

 （2）

当 （3）

是关于z =（z1，z2，...，zD）的信息量，它是x的等价向量，；

 （4）

是关于概率向量p =的信息量（p1，p2，...，pD）

 （5）

pd（0,1），;and > 0是增强因子。方程式中的第一项 （2）是衡量信息丢失后的将原始矢量x变换为概率矢量p。因为xdR，d = 1,2，。。 。，D，我们无法直接获取x的信息量。 因此，执行线性变换在原始向量x上，然后生成其等效向量z如:

 (6)

其中xmin = min {x1，x2，。。 。，xD}。 对于原始向量 =（x1，x2，...，xD），k{1,2，。。 。，D}，xk0，Eq。 （6）确保0 <zd <1 for€{1,2，。。 。，D}。 方程的作用。（6）是方便的信息量的计算。我们无法计算信息 - 原始向量的数量直接如果元素原始向量超出区间（0,1）。 因此，我们需要将原始矢量转换为等效矢量元素都在区间（0,1）内。 第二个任期方程（2）是控制概率向量元素之间的多样性p1，p2，。。 。，pD。 Hp达到其最大值In（D）at p1 = p2 =···= pD = 我们不仅希望最佳温度参数最小化变换的信息损失，但也最大化概率向量元素之间的多样性。 因此，我们可以得到的最优表达式:

 (7)

让 并且  ，Hp in Eq

1. 可以写成：

 （8）

将（8）代入（2）可以得到

 （9）

通过解决确定的分析公式非常困难.因为E和F是相关的函数

到，我们试图找到最小的E或F，它可以最小化L（T）in式（9）通过计算

 （10）

或者

 （11）

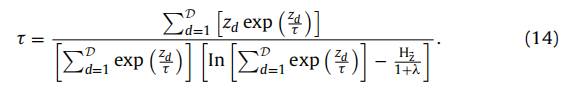
因为E>0,F>0 因此我们有

 （12）

即

 （13）

1. 可以简化为：



式。 （14）是Opti-Softmax方法的启发式更新规则确定最佳温度参数。 根据

此更新规则，Opti-Softmax方法如算法1所示 给出一个迭代过程来确定.

算法1：Opti-Softmax method

1输入：原始向量x =（x1，x2，...，xD），xdR，d = 1,2，。。 。，D

和{1,2，。。 。，D}，xk 0; 增强因子> 0;该停止阈值> 0; 初始温度参数> 0。

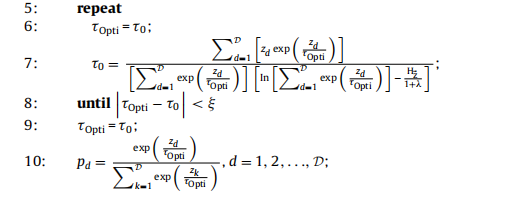
2输出：输出：概率向量p =（p1，p2，...，pD），pd（0,1），d = 1,2，。。 。，D，

和最佳温度参数

3计算等效矢量z =（z1，z2，...，zD），zd（0,1） d= 1,2，。。 。，D，

根据方程式（6）

4 根据方程式计算z的信息量Hz。（3）;



**Experimental simulations**实验模拟

进行了两个实验以证明其可行性和Opti-Softmax方法的有效性。 第一个实验是

显示更新规则，如公式1所示。 （14）是收敛的，第二个实验是用Opti-Softmax方法处理

强化学习中的D-armed bandit问题。

在第一个实验中，一个10维原始矢量x =（ - 7013.7933，-7282.7121,649.9646,4515.7862，-2025.9390，-2831.6297，-4294.4118,7372.7049,2528.2535，-5176.5538）是在区间[-10000,10000]中随机生成。 我们测试了Opti-Softmax方法的工作性能（增强因子= 1和停止阈值=，不同的初始温度参数0.001和1.实验结果列于图2和3中。

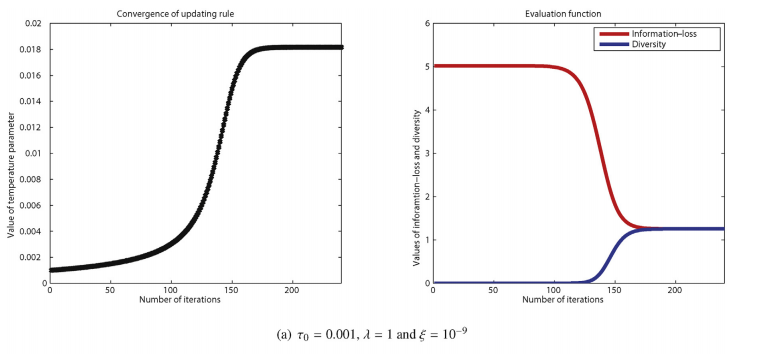
对于不同的初始温度参数（ = 0.001 并且 = 1），更新规则 （14）制作Opti-Softmax方法收敛到相同的最佳温度参数 Opti = 0.0184。 图2（a）和（b）的左侧子图显示了这一点更新曲线逐渐从0.001增加到0.0184240次迭代并从95逐渐减少到0.0184迭代， Opti-Softmax方法可以在不依赖初始值的情况下找到最佳温度参数。

（14）带来的收敛即 信息丢失术语）和（即，多样性项），如图所示

在图2（a）和（b）的右边的子图中。 同时，随着温度参数增加，逐渐减小（即，信息损失逐渐减少），而逐渐增加（即多样性逐渐减少）。

增强因子影响了物质的数量选择最佳温度参数。 让我们将增强因子，每次增加0.05从0到1.5中。 迭代次数和最优温度参数对应不同的初始值温度参数 = 0.001和 = 1总结于图3.我们可以看到迭代次数首先减少，然后随着增强因子的增加而增加如图3（a）和（b）所示。 在图3（c）中，最佳温度参数随着增加而逐渐减小。当> 1时，迭代次数显示减少的趋势。 这为我们提供了选择适当增强的启示

因素，因为较小的Opti和更大的结果导致更大的信息损失。 通常，我们选择增强因子在区间（0,1）当等价向量如被使用的公式6所示。

图2

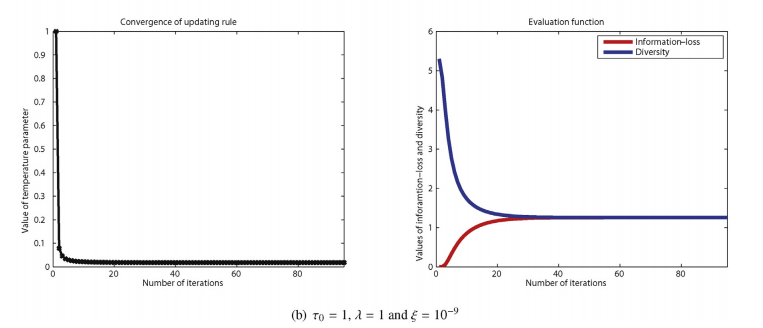
 图2

图2.最佳温度参数Opti在原始矢量上为0.0184（= =（-3.7933，-2.7121,649.9646,4515.7862，-5.9390，-1.6297，-4.4118，7372.7049,2528.2535，-6.5538）和相应的概率向量p为（0.0017,0.0015,0.0403,0.176,6.0134,0.0096,0.0053,0.6394,0.0873,0.0037）。

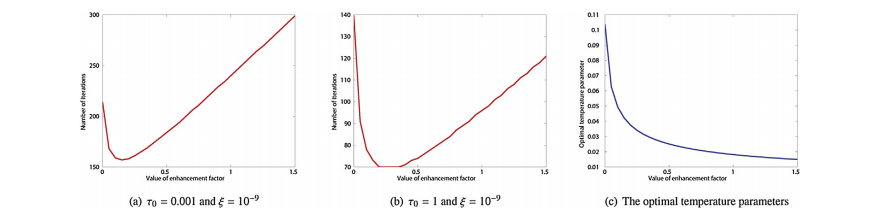


图3

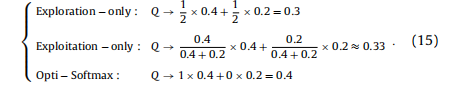
图3.增强因子对迭代次数和最佳温度参数的影响，其中原始向量x为（-3.7933，-7282.7121，

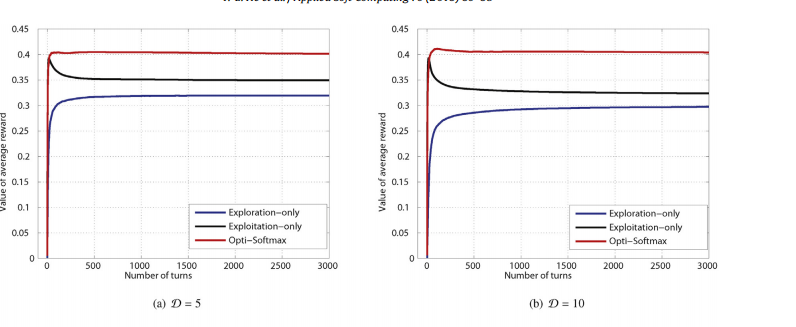
649.9646,4515.7862， -2025.9390， -2831.6297，-4294.4118,7372.7049,2528.2535，-5176.5538）。

与仅探索和仅剥削相比方法，我们验证Opti-Softmax的实际性能处理D-armed bandit问题时的方法（D =5,10,20和50）在强化学习中。 有两种类型在这个实验中的强盗臂：奇数bandit-arm以概率0.4和偶数返回奖励1bandit-arm以0.2的概率返回奖励1。 游戏玩法转为D = 5,10和D = 20,50分别为3000和5000，

分别。 对于每种方法，我们重复游戏100次记录赌徒的平均奖励。 实验结果是

如图4所示，其中Opti-Softmax方法的参数设置为 = 1，= 1和在图4中，我们可以看到Opti Softmax方法获得的性能明显优于仅探索和仅开发的方法。 平均Opti-Softmax方法的奖励近似为0.4。赌徒在当前实验环境中可以获得的最高奖励。 设N表示游戏次数。 当仅探索Q的平均奖励时，仅开发和Opti-Softmax方法计算如下：



 图4

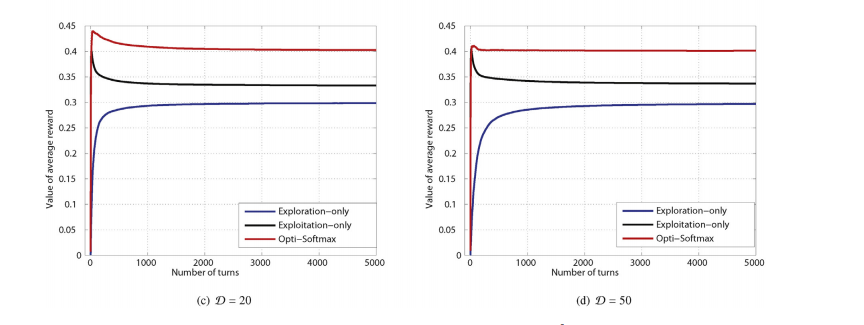


图4

图4.仅探索，仅开发和Opti-Softmax的比较（ = 1，= 1和解决D-armed bandit问题

式。（15）反映bandit-arms使得赌徒获得以概率选择概率为0.4的奖励1在Opti-Softmax方法中。 为了证实这一结论，我们进行了如图所示的补充实验图5.对于20个武装的强盗问题，我们计算每个强盗臂在5000脳100匝中被选择的数量。 我们很容易就能发现奇数的总数

强盗臂由Opti-Softmax方法选择（红色条）近似于493812，即选择奇数编号的匪臂的概率是=0.9876≤1。 对于仅探索和仅限于开采的方法，选择奇数编号的武器臂的概率是=0.4992≤1（绿色条）和=0.6599≤（蓝色条）。

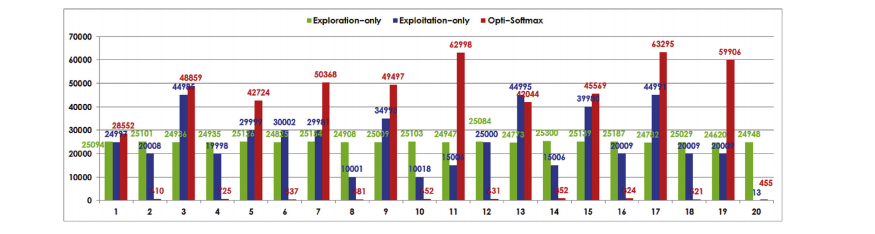


图5

图5.选择20个武装匪徒问题中每个强盗臂的数量。 （为了解释文本中对颜色的引用，读者可以参考网络这篇文章的版本。）

**Conclusions and further works**

通过设计有效的评估功能和更新规则，本文提出了一个有用而简单的方法

Opti-Softmax确定最佳温度参数用于强化学习中的Softmax函数。 实验结果表明，Opti-Softmax方法是可行的有效的，不仅可以找到Softmax函数的最优温度参数，而且迭代次数较少，而且可以得到赌徒在玩D型强盗时获得更高的奖励游戏。 事实上，Opti-Softmax方法是一种基于不确定性约简的参数优化技术。 在未来的作品中，我们将研究Opti-Softmax方法与不确定性的整合基于减少的机器学习方法，例如随机权重网络[7,9,10]，代表性学习[3,20]，不完整的信息处理[2,14]。