第 44 卷 第 11 期 计 算 机 工 程 2018 年 11 月
Vol. 44 No. 11 Computer Engineering November 2018

·人工智能及识别技术· 文章编号：1000-3428(2018)11-0172-06 文献标志码：A 中图分类号：TP181

针对新用户冷启动问题的改进 Epsilon-greedy 算法

王素琴[1]，张 洋[1]，蒋 浩[2]，朱登明[2]

（1. 华北电力大学 控制与计算机工程学院，北京 102206；2. 中国科学院计算技术研究所，北京 100080）

摘 要： 在解决新用户冷启动问题时，固定不变的 Epsilon 参数会使传统 Epsilon-greedy 算法收敛缓慢。为此，提出一种改进的 Epsilon-greedy 算法。利用免疫反馈模型动态调整 Epsilon 参数，从而使算法快速收敛。使用蒙特卡罗模拟方法对算法进行实验验证，结果表明，该算法能够在用户与推荐系统交互较少的情况下为用户进行有效推荐，且推荐效果优于传统的 Epsilon-greedy、Softmax 和 UCB 算法。

关键词： 推荐系统；冷启动；Epsilon-greedy 算法；免疫反馈模型；bandit 算法

中文引用格式： 王素琴，张 洋，蒋 浩，等. 针对新用户冷启动问题的改进 Epsilon-greedy 算法[J]. 计算机工程，2018，44(11)：172-177.

英文引用格式： WANG Suqin，ZHANG Yang，JIANG Hao，et al. Improved Epsilon-greedy algorithm for cold-start problem of new users[J]. Computer Engineering，2018，44(11)：172-177.

Improved Epsilon-greedy Algorithm for Cold-start Problem of New Users

WANG Suqin[1]，ZHANG Yang[1]，JIANG Hao[2]，ZHU Dengming[2]

（1. School of Control and Computer Engineering，North China Electric Power University，Beijing 102206，China；
2. Institute of Computing Technology，Chinese Academy of Sciences，Beijing 100080，China）

【Abstract】 When solving the cold-start problem of new users，fixed and invariant Epsilon parameters will slow the convergence of traditional Epsilon-greedy algorithm. Therefore，an improved Epsilon-greedy algorithm is proposed. Immune feedback model is used to dynamically adjust the Epsilon parameters so that the algorithm converges quickly. Monte Carlo simulation is used to validate the proposed algorithm. Results show that this algorithm can effectively recommend to users when they have little interaction with the recommendation system，and the recommendation effect is better than the traditional Epsilon-greedy algorithm，Softmax algorithm and UCB algorithm.

【Key words】 recommendation system；cold-start；Epsilon-greedy algorithm；immune feedback model；bandit algorithm

DOI： 10.19678/j.issn.1000-3428.0048631

0 概述

电子商务和社交媒体网站通常都面临着严重的信息过载问题，推荐算法是解决该问题的有效手段之一。推荐算法一般根据用户与网站的交互记录来推测用户可能喜欢的物品或者友人（以下统称为物品）。当新用户登录到系统中时，由于他们没有或者只有少量购买记录和浏览记录，推荐系统很难为他们进行有效的推荐，该问题被称为新用户冷启动问题[1]。

快速发展中的电子商务网站或者社交媒体网站，吸引着大量新用户的涌入，但由于存在新用户冷启动问题，不能为他们提供充分的有用信息，常常导致新用户大量流失。

为解决上述问题，有研究者提出推荐算法。传统的推荐算法有协同过滤算法[2-3]、基于内容的推荐算法[4]以及混合推荐算法[5]等。协同过滤算法基于用户的行为记录，计算用户之间的相似度或者物品之间的相似度并进行推荐。但由于新用户没有或者只有少量的行为记录，导致协同过滤算法难以为其进行有效推荐。基于内容的推荐算法首先提取用户或者物品的特征，根据用户或者物品之间特征的相似度来进行推荐。但新用户可能没有任何特征或者只有少量特征可以提取，因此，基于内容的推荐算法也难以取得良好效果。混合推荐算法是指将多种推荐算法有机地结合起来进行推荐，但实际应用结果表明，混合推荐算法也不能很好地解决新用户冷启动问题。

基金项目： 国家自然科学基金“逼真稳定的服装动画方法研究”(61300131)；北京市共建项目(2014JG48)。

作者简介： 王素琴(1970—)，女，副教授、硕士，主研方向为数据挖掘、计算机视觉；张 洋，硕士；蒋 浩，助理研究员、博士；朱登明，副研究员、博士。

收稿日期： 2017-09-11 **修回日期：** 2017-10-28 **E-mail：** wsq@ncepu.edu.cn

目前,有很多学者对新用户冷启动问题进行了深入研究,提出了若干解决该问题的算法。其中,最简单的方法是随机推荐,当新用户进入系统时,推荐算法在物品库中随机选择若干个物品推荐给用户,但这种推荐结果往往无法令用户满意。文献[6-7]提出基于偏好的推荐算法,在所有用户中查找与当前用户偏好相近的用户,根据领域相关度、评价相似度等对相似用户进行筛选,得到与当前用户相似度最高的一批用户,根据这批用户的偏好信息为当前用户进行推荐。该算法相对于协同过滤算法在冷启动问题上有一定改进,但是仍不能解决完全没有任何信息的新用户冷启动问题。文献[8]提出对已有的用户信息或物品信息进行分类,当新用户或新物品进入系统时利用贝叶斯分类方法将其归类到相应的用户类或物品类。这种方法在系统已经获取用户的基础信息或者少量的用户交互信息时能给出较好的推荐,但是当一个完全没有任何信息的新用户登录到系统时,这类方法不能取得较好效果。文献[9]基于用户的人口统计学信息为新用户进行推荐,但其在未获取新用户的人口统计学信息时无法使用。

以上各种解决新用户冷启动问题的算法都是在已知用户信息或者用户与推荐系统有过一些交互的基础上进行的推荐,这些算法在未获知用户信息、用户与系统没有交互或者交互非常少的情况下难以取得令人满意的效果。为解决该问题,本文建立 N 臂老虎机(N-armed bandit)[10] 和免疫反馈(Immune Feedback)[11] 相结合的模型,提出一种改进的 Epsilon-greedy 算法 EGIF 为新用户进行推荐,并根据用户的反馈及时调整策略,以不断改善算法的推荐效果[12]。

1　N 臂老虎机模型

一台老虎机有多个拉杆,拉动每个拉杆可能获得 0 个或 1 个金币,这个回报是随机的。如果拉动拉杆的次数有限,那么如何在老虎机上获得最大回报(即最多的金币数),就是所谓的 N 臂老虎机问题。

为了获得最大回报,赌徒应该尽快找出回报率最高的拉杆。最简单的方法是给每个拉杆相同的尝试次数(如 10 次),统计每个拉杆返回的金币个数,找到返回金币数最多的拉杆,以后一直选择这个拉杆以期获得最大回报。但这个策略存在一定缺陷:1)回报率低的老虎机可能在 10 次内返回比回报率高的拉杆更多的金币;2)如果拉杆较多,给每个拉杆 10 次实验机会无疑会造成很大的浪费。

目前,解决 N 臂老虎机问题的算法大致分 3 类:

1)无指导“探索”算法,如 Epsilon-greedy[13]、Epoch-greedy[14] 等,这类算法每次以 Epsilon 的概率在 N 臂老虎机中随机选择一个拉杆,否则就选择目前为止平均收益最大的那个拉杆。

2)有指导“探索”算法,如 UCB(Upper Confidence Bound)[15]、EXP4[16] 等,这类算法每次选择置信上限最高的那个拉杆。

3)概率匹配算法,如 Thompson sampling[17] 等,这类算法假设每个拉杆产生收益的概率 p 符合 beta 分布,通过实验不断调整 beta 分布的参数,每次选择拉杆的方式为:用每个拉杆现有的 beta 分布产生一个随机数,选择所产生的随机数最大的那个拉杆。

在这些解决 N 臂老虎机问题的算法中,Epsilon-greedy 算法和 UCB 算法在实际应用中表现较优秀。

1.1　Epsilon-greedy 算法

Epsilon-greedy 算法是一种解决 N 臂老虎机问题的简单算法。greedy 算法总是选择在当前时刻算法认为最好的动作。Epsilon-greedy 算法和 greedy 算法非常相似,它一般会选择最好的动作,也可能去“探索”其他可行的动作。Epsilon-greedy 算法每次以 Epsilon 的概率去“探索”(在所有拉杆中随机选择一个拉杆),以 1 – Epsilon 的概率去“发现”(选择之前“探索”到的回报率最高的那个拉杆)。

在 Epsilon-greedy 算法中,比较关键的问题是如何确定 Epsilon 的值。如果 Epsilon 的值较大,会增加探索的概率,这虽能够加快算法的收敛,但往往不能很好地利用已经探索到的成果,导致结果较差;如果 Epsilon 的值较小,模型的稳定性更好,但会使算法的收敛速度降低。

1.2　UCB 算法

UCB 算法在每轮选择置信上限 $J(t)$ 最大的拉杆,$J(t)$ 计算公式如下:

$$J(t)=\underset{i=1,2,\cdots,k}{\operatorname{argmax}}\left(\mu_i+2\sqrt{\frac{2\ln t}{n_i}}\right) \tag{1}$$

$$\mu_i=\frac{c(g)}{c(t)} \tag{2}$$

其中,μ_i 为实际观测到的老虎机返回金币的概率,$c(g)$ 为返回金币个数,$c(t)$ 为全部尝试次数,n_i 为 t 轮内尝试拉动第 i 个拉杆的次数。

UCB 算法将“探索”和“发现”2 个过程融合到一个公式中,能够让拉动次数较少的拉杆有更多被尝试的机会。

1.3　N 臂老虎机模型在推荐算法中的应用

N 臂老虎机问题是优化一个同时玩多个老虎机的赌徒的收入统计问题[18-19]。文献[20]基于内容的 N 臂老虎机模型提出 LinRel 算法,文献[21]提出 LinUCB 算法,改进了 LinRel 算法并用基于内容的 N 臂老虎机模型在新闻文本推荐中对用户反馈进行建模。N 臂老虎机模型应用于推荐算法时,将 N 臂老虎机的拉杆定义为将要推荐给用户的物品,将拉动拉杆的动作定义为给用户进行推荐,将拉杆返回的奖励定义为用户点击了为其推荐的物品。

在为用户推荐物品时,每次为用户推荐一个物品后收集用户对此物品的评分。根据已知的用户评分决定下一轮应该推荐的物品。此时,可以继续推荐用户之前评分最高的物品,也可以随机选择一个物品推荐给用户,前者称为“发现”,后者称为“探索”。

N 臂老虎机模型能够根据用户的在线反馈为新用户进行合理的推荐,同时由于老虎机模型中存在的“探索”部分,能够提高推荐结果的多样性[22]。

2 免疫反馈模型

免疫系统是人类和脊椎动物所拥有的防御系统,是由许多执行免疫功能的免疫器官、免疫细胞和免疫分子等组成的复杂自适应系统。免疫系统使机体免受病原体、有害物质以及癌细胞等致病因子的侵害。在免疫系统中,有一种免疫反馈机制同时完成 2 项任务:1)对出现的抗原快速反应;2)使免疫系统快速达到稳定平衡。

免疫反馈机制的原理如图 1 所示,人体的免疫细胞大致分为 T 细胞和 B 细胞 2 种。T 细胞的主要功能是吞噬入侵的抗原,B 细胞的主要功能是产生多种抗体,用抗体来中和入侵的抗原。当抗原物质(细菌、病毒)入侵人体后,会同时刺激辅助 T 细胞和抑制 T 细胞。一方面,辅助 T 细胞能够协助 B 细胞产生抗体,促进 Killer T 细胞的生成;另一方面,抑制 T 细胞也会抑制 B 细胞产生抗体,抑制 Killer T 细胞的生成。通过免疫系统中辅助 T 细胞和抑制 T 细胞之间的相互作用,使免疫系统实时的处在抗原和抗体的动态平衡中。这种动态平衡使得免疫系统能够对入侵的抗原做出快速反应,并且能使免疫系统迅速达到平衡。

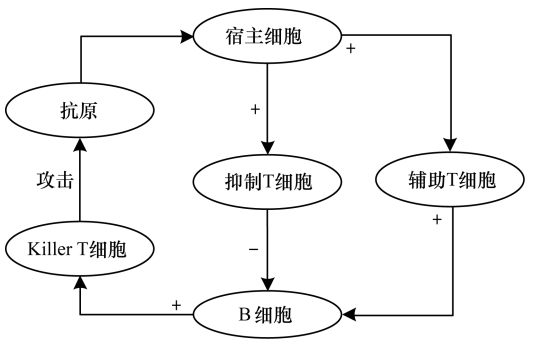

图 1 免疫反馈机制原理

研究人员借鉴免疫反馈思想来加速神经网络算法的收敛速度[23-24]。本文将免疫反馈模型应用于 Epsilon-greedy 算法,使 Epsilon-greedy 算法能够更快收敛,从而更好地为新用户进行推荐。

辅助 T 细胞对 B 细胞的刺激定义为:

$$T_{\text{help}}(k)=K_1\varepsilon(k) \tag{3}$$

其中,K_1 为 T 细胞对 B 细胞的刺激程度,$\varepsilon(k)$ 为第 k 代B 细胞的数量。

抑制 T 细胞对 B 细胞的抑制定义为:

$$T_{\text{sup}}(k)=K_2\{T_{\text{kill}}(k-d)-T_{\text{kill}}(k-d-1)\}^2\varepsilon(k) \tag{4}$$

其中,K_2 为抑制因子,d 为超参数,表示第 d 代,$T_{\text{kill}}(k-d)$ 和 $T_{\text{kill}}(k-d-1)$ 分别为第 $k-d$ 代和第 $k-d-1$ 代 Killer T 细胞数量。

Killer T 细胞接受到的总刺激为:

$$T_{\text{Killer}}(k)=T_{\text{help}}(k)-T_{\text{sup}}(k) \tag{5}$$

若以 $\Delta reward$ 作为抗原,Epsilon-greedy 算法中的 $Epsilon$ 参数值作为抗体,则有如下关系式[21]:

$$Epsilon(k)=K_p[1-\gamma\{Epsilon(k-d)-Epsilon(k-d-1)\}^2]\Delta reward \tag{6}$$

其中,$K_p=K_1$,$\gamma=K_2/K_1$。参数 K_p 控制免疫系统对抗原的反应速度,参数 γ 控制免疫系统的稳定性。$Epsilon(k)$ 为第 k 次对用户推荐时 $Epsilon$ 的值,$reward$ 为用户对推荐物品的喜好程度,其计算参考式(7)。

$$reward(k)=\frac{click(k)}{recommend(k)} \tag{7}$$

其中,$click(k)$ 是用户的点击率,$recommended(k)$ 是为用户推荐物品的总数。$\Delta reward$ 的计算如式(8)。

$$\Delta reward=reward(k)-reward(k-d) \tag{8}$$

3 EGIF 算法

本文提出的 EGIF 算法将 Epsilon-greedy 算法和免疫反馈模型相结合,用以解决新用户冷启动问题。

将给用户推荐的 N 个物品定义为 N 臂老虎机的 N 个拉杆,每次为用户进行推荐相当于一次拉动老虎机拉杆的动作,每次推荐后用户是否点击了所推荐的物品相当于用户在拉动某一个拉杆时是否获得了奖励。

Epsilon-greedy 算法以 Epsilon 的概率去随机“探索”用户的潜在偏好,并为用户推荐物品,以 1-Epsilon 的概率去利用已经“探索”到的用户偏好来为用户推荐物品。传统的 Epsilon-greedy 算法的 Epsilon 参数值是固定不变的,如果 Epsilon 取值较小,算法在短时间内不容易“探索”到用户的潜在兴趣,导致算法的收敛速度较慢,随着时间的推移,在“探索”到用户的兴趣后,能以很大的概率去利用已经“探索”到的用户兴趣并进行推荐,从而取得较好的推荐效果。如果 Epsilon 取值较大,虽然算法能够更快的收敛,在较短时间内“探索”到用户的兴趣,但是在“探索”到用户兴趣后仍然保持很大的概率去“探索”,而不是根据已经“探索”到的用户兴趣进行推荐,这会导致算法的推荐效果较差。不同 Epsilon 值的平均推荐回报率如图 2 所示。

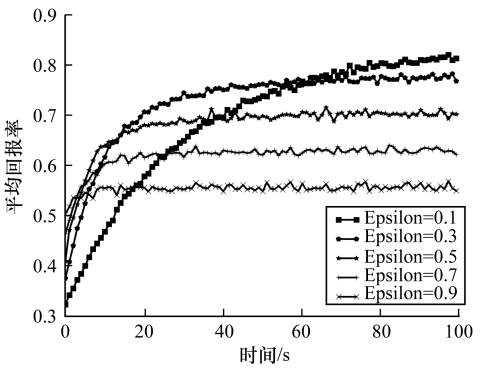

图 2　不同 Epsilon 值的平均回报率

在免疫反馈系统中,当抗原入侵机体后,在抗原的刺激下抗体数量迅速上升。随着抗体数量的增多,抗体会抑制自身的增长使抗体数量迅速下降,以便免疫系统保持平衡。本文利用免疫反馈模型来动态调整 Epsilon 参数的值,使算法既能很快收敛,又能实现很好的推荐效果。根据式(6),在用户刚进入系统时,Epsilon 的值会迅速升高,以尽快“探索”用户的偏好。随着用户与系统交互次数的增多,Epsilon 的值会迅速降低,以便更好地利用已“探索”到的用户偏好进行推荐。本文 EGIF 算法具体描述如下:

```
1. begin
2. 初始化拉杆的回报分布为用户随机推荐物品
3. k = 0
4. while true do
5. begin
6. if 用户点击了推荐物品 then
7. 返回 1
8. else 返回 0
9. 记录推荐物品、用户是否点击物品及点击次数
10. 根据式(7)计算 reward(k)
11. 根据式(8)计算 Δreward
12. 根据式(6)计算 Epsilon
13. r←0 到 1 之间的随机数
14. if r > Epsilon then
15. 随机推荐物品
16. else 推荐点击次数最多的物品
17. 更新拉杆的回报分布
18. k + +
19. end
20. end
```

4　实验结果及分析

EGIF 算法需要实时分析和选择将要推荐给用户的物品,这意味着算法的行为要依赖于其看到的数据,算法所看到的数据又依赖于算法的行为。算法数据和算法行为的关系如图 3 所示。

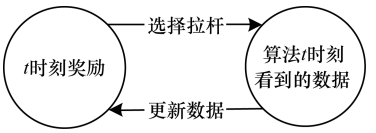

图 3　算法数据与算法行为间的关系

因此,EGIF 算法是一种在线算法,算法的测试需要在真实的系统中根据用户反馈来进行,但是在真实系统中进行测试的风险很高。为解决这一问题,本文采用蒙特卡罗模拟方法[25]进行测试。该方法能够提供实时的模拟数据供算法分析,从而对算法进行评测。

本文的实验环境为:Intel Core i5 CPU,2.7 GHz 主频,Windows 7 操作系统,4 GB 内存,编程语言为 Python。在实验中,N 臂老虎机的拉杆为伯努利拉杆,共设置 5 个不同的拉杆,每个拉杆返回奖励的概率分别为 0.1、0.2、0.3、0.5、0.9。实验共进行 5 000 个epoch 模拟,每个 epoch 按顺序拉动 500 次拉杆,记录每个拉杆返回的 reward 值以及总 reward 值。

将 EGIF 算法应用于推荐系统,令用户点击算法推荐的物品的概率为 p,没有点击算法推荐的物品的概率为 $1-p$。每个拉杆的回报率各不相同,例如,拉杆 1 的回报率为 0.1,代表拉动拉杆 10 次,会有一次返回的奖励为 1,其余时候返回的奖励为 0。回报率表示推荐系统中用户对每件物品的偏好程度。

在实验中,对 EGIF 算法进行测试并返回一个测试结果数据集,以说明每次模拟时算法选择了哪个拉杆以及算法在每一个时间点的表现。由于每次模拟都根据随机数生成,实验结果的噪声很大,因此需要进行多次模拟。EGIF 算法和固定 Epsilon 的 Epsilon-greedy 算法的平均奖励指标实验结果如图 4 所示,其中,Epsilon-greedy 算法的 Epsilon 超参数值分别选取 0.1、0.3、0.5、0.7 和 0.9。从图 4 可以看出,用户与系统交互约 5 次时,EGIF 算法在平均奖励上已经超出了固定 Epsilon 的 Epsilon-greedy 算法,表明 EGIF 算法能够在极短的时间内找到用户的偏好,为用户进行更好的推荐。从图 4 还能看出,不仅是在极短的时间内 EGIF 算法能够找到用户的偏好,在以后更长的时间里,EGIF 算法能够维持在 0.85的平均奖励,而固定 Epsilon 的 Epsilon-greedy 算法在与用户交互 100 次时也只能达到 0.8 的平均奖励。综上,与固定 Epsilon 的 Epsilon-greedy 算法相比,EGIF 算法不仅能更好地解决新用户冷启动问题,而且能够在较长的时期内维持较佳的推荐效果。

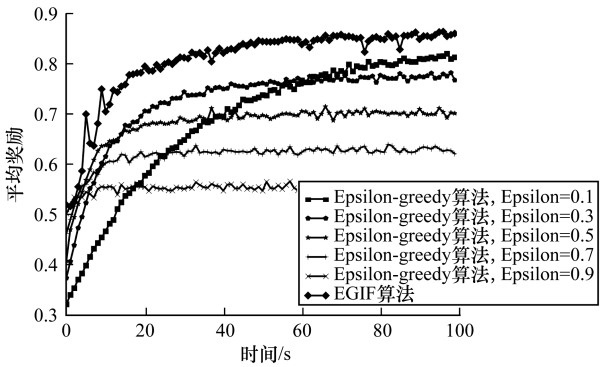

图 4 2 种算法平均奖励比较

图 5、图 6 所示分别为 EGIF 算法和固定 Epsilon 的 Epsilon-greedy 算法的奖励总和以及选择到最好拉杆的概率结果。从图 5、图 6 可以看出，相对于固定 Epsilon 的 Epsilon-greedy 算法，本文 EGIF 算法性能更优。

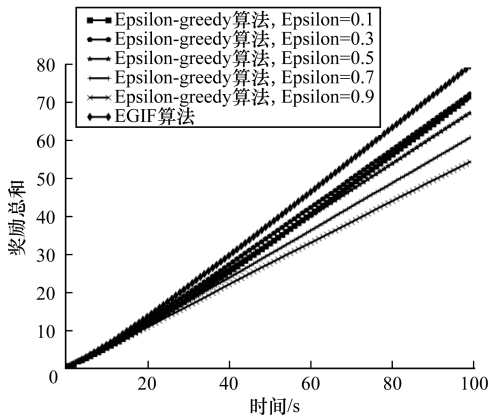

图 5 2 种算法奖励总和比较

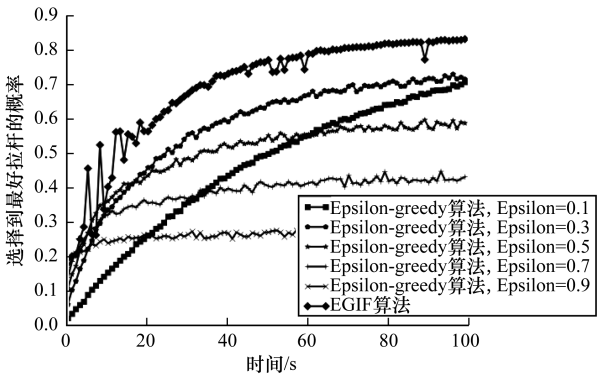

图 6 2 种算法选择到最好拉杆的概率比较

EGIF 算法在平均奖励指标上与 Softmax 算法、Annelealing 算法以及 UCB 算法的比较如图 7 所示。从图 7 可以看出，EGIF 算法在用户与系统交互的前 20 次中，虽然平均奖励值有所波动，但是该值几乎一直高于 Softmax 算法和 Anneleanling 算法，而 UCB 算法波动剧烈，平均奖励结果非常不稳定。EGIF 算法在平均奖励方面之所以有一定的波动，是因为免疫反馈模型会根据算法平均奖励的变化动态地调整 Epsilon 参数的值，以使算法能够快速收敛。

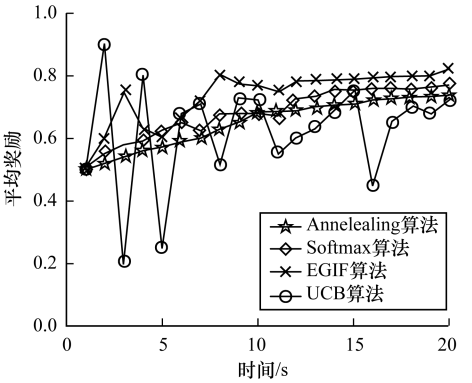

图 7 4 种算法平均奖励对比

图 8、图 9 所示分别为 EGIF 算法、Softmax 算法、Annelealing 算法以及 UCB 算法在奖励总和指标和选择到最好拉杆概率指标上的比较结果。从图 8、图 9 可以看出，EGIF 算法在这 2 个指标上的表现都优于对比算法。

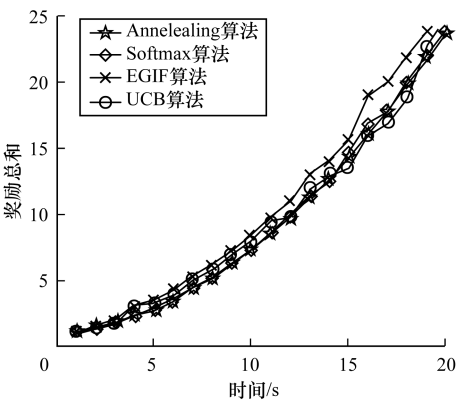

图 8 4 种算法奖励总和对比

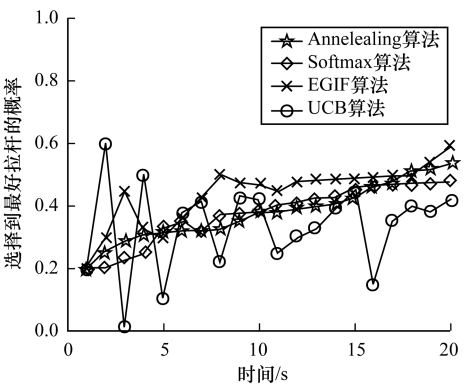

图 9 4 种算法选择到最好拉杆的概率对比

5　结束语

本文将 Epsilon-greedy 算法和免疫反馈模型相结合,提出一种解决新用户冷启动问题的 EGIF 算法。在传统的 Epsilon-greedy 算法中利用免疫反馈模型动态调整 Epsilon 的参数值,以解决传统 Epsilon-greedy 算法不能快速收敛的问题。实验结果表明,EGIF 算法能够在新用户进入系统时迅速找到用户的偏好,为新用户进行更好的推荐。在解决 N 臂老虎机问题的诸多算法中,除 Epsilon-greedy 算法外,其他算法也存在“探索”和“发现”问题,本文仅对 Epsilon-greedy 算法和免疫反馈模型相结合进行了研究。下一步考虑将免疫反馈模型与其他算法相结合,以进一步提高本文推荐算法的性能。

参考文献

[1] LIU C,WANG Y. Analysis on the cold-start problem in recommendation system[J]. Telecommunications Network Technology,2017(1):56-76.

[2] WEI J,HE J,CHEN K,et al. Collaborative filtering and deep learning based recommendation system for cold start items[J]. Expert Systems with Applications,2017,69:29-39.

[3] 冷亚军,陆　青,梁昌勇. 协同过滤推荐技术综述[J]. 模式识别与人工智能,2014,27(8):720-734.

[4] 王　洁,汤小春. 基于社区网络内容的个性化推荐算法研究[J]. 计算机应用研究,2011,28(4):1248-1250.

[5] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.

[6] ZHU R,WANG H M,FENG D W. Trustworthy services selection based on preference recommendation[J]. Journal of Software,2011,22(5):852-864.

[7] 李　改,李　磊. 一种解决协同过滤系统冷启动问题的新算法[J]. 山东大学学报(工学版),2012,42(2):11-17.

[8] MASSA P, AVESANI P. Trust-aware recommender systems[C]//Proceedings of ACM Conference on Recommender Systems. New York, USA: ACM Press, 2007:17-24.

[9] MIDDLETON S E,SHADBOLT N R,DE ROURE D C. Ontological user profiling in recommender systems[J]. ACM Transactions on Information Systems,2004,22(1):54-88.

[10] ZHANG X,NAKHAI M R,WAN N S F W A. A multi-armed bandit approach to distributed robust beamforming in multicell networks[C]//Proceedings of 2016 IEEE Global Communications Conference. Washington D. C., USA:IEEE Press,2016:1-6.

[11] TAKAHASHI K,YAMADA T. Application of an immune feedback mechanism to control systems[J]. JSME International Journal,1998,41(2):184-191.

[12] BERRY D A, FRISTEDT B. Bandit problems: sequential allocation of experiments[M]. Berlin, Germany: Springer, 1985.

[13] HILLS T T. Trade-off between exploration and exploitation[M]//TODD K, SHACKELFOR D, VIVIANA A. Encyclopedia of evolutionary psychological science. Berlin, Germany: Springer, 2017.

[14] LANGFORD J,ZHANG T. The epoch-greedy algorithm for contextual multi-armed bandits[EB/OL]. [2017-09-01]. http://courses. cms. caltech. edu/cs101. 2/slides/cs101. 2-05-contextual-bandits. pdf.

[15] AUER P, ORTNER R. UCB revisited: improved regret bounds for the stochastic multi-armed bandit problem[J]. Periodica Mathematica Hungarica,2010,61(1):55-65.

[16] AUER P, CESA-BIANCHI N, FREUND Y, et al. The nonstochastic multiarmed bandit problem[J]. SIAM Journal on Computing,2002,32(1):48-77.

[17] CHAPELLE O, LI L. An empirical evaluation of thompson sampling[C]//Proceedings of International Conference on Neural Information Processing Systems. [S. l.]:Curran Associates Inc.,2011:2249-2257.

[18] BERRY D A, FRISTEDT B. Bandit problems[J]. Monographs on Statistics and Applied Probability,1985,25(10):1585-1594.

[19] ANANTHARAM V, VARAIYA P, WALRAND J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part II: Markovian rewards[J]. IEEE Transactions on Automatic Control,1987,32(11):977-982.

[20] AUER P. Using confidence bounds for exploitation-exploration trade-offs[J]. Journal of Machine Learning Research,2002,3(3):397-422.

[21] LI L,CHU W,LANGFORD J,et al. A contextual-bandit approach to personalized news article recommendation[C]// Proceedings of International Conference on World Wide Web. New York, USA: ACM Press, 2010: 661-670.

[22] MCNEE S M,RIEDL J,KONSTAN J A. Being accurate is not enough: how accuracy metrics have hurt recommender systems[C]//Proceedings of 2006 Conference on Human Factors in Computing Systems. New York,USA:ACM Press,2006:1097-1101.

[23] KAWAFUKU M, SASAKI M, TAKAHASHI K. Adaptive learning method of neural network controller using an immune feedback law[C]//Proceedings of IEEE/ASME International Conference on Advanced Intelligent Mechatronics. Washington D. C., USA: IEEE Press, 1999: 641-646.

[24] SASAKI M, KAWAFUKU M, TAKAHASHI K. An immune feedback mechanism based adaptive learning of neural network controller[C]//Proceedings of International Conference on Neural Information Processing. Washington D. C., USA:IEEE Press,1999:502-507.

[25] MACKAY D J C. Introduction to Monte Carlo methods[C]// Proceedings of NATO Advanced Study Institute on Learning in Graphical Models. Norwell,USA:Kluwer Academic Publishers, 1998:175-204.

编辑　吴云芳